DEVELOPING TRUST AND MANAGING UNCERTAINTY IN PARTIALLY OBSERVABLE SEQUENTIAL DECISION-MAKING ENVIRONMENTS

A Dissertation
Presented to
The Academic Faculty

By

R. Reid Bishop

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial & Systems Engineering

Georgia Institute of Technology

December 2019

DEVELOPING TRUST AND MANAGING UNCERTAINTY IN PARTIALLY OBSERVABLE SEQUENTIAL DECISION-MAKING ENVIRONMENTS

Approved by:

Prof. Chelsea C. White III, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Prof. Enlu Zhou
School of Industrial and Systems Engineering
Georgia Institute of Technology

Prof. Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Prof. He Wang
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Brandon Eames
Sandia National Laboratories

Dr. Alexander Outkin
Sandia National Laboratories

Date Approved: October 7, 2019

To my parents, who love me and encourage me always.

ACKNOWLEDGEMENTS

I feel very blessed to have so many people in my life to whom I owe immense gratitude for their unwavering support during my PhD studies, and without whom I cannot imagine completing this dissertation.

First, I would like to thank my doctoral advisor, Prof. Chip White. In my time at Georgia Tech, I have grown in so many ways, and I owe much of that maturity — in research, in problem-solving, in systematic thinking — to the consistent guidance and mentoring that Chip freely provided. As I approach the next phase of my life and career, I am keenly aware that my regular discussions with Chip were the best educational and developmental experiences that I have ever had. Moreover, and even more importantly, I always felt that Chip genuinely cared for my holistic personal development, and not just my research ideas and output. In early 2017, when I endured health struggles, Chip did all he could do to ensure that I had the time and space to both heal and continue my studies. That period of time led to many of the ideas that compose this dissertation. For that, I will always be grateful.

Secondly, I have benefited in many ways from the research relationship that I have had with Sandia National Laboratories on the Practical Evaluation and Synthesis of Trust in Government Systems (PRESTIGE) project. On one level, their funding allowed me to pursue the research in this dissertation apart from financial insecurity. On another level, my thinking about the research in this dissertation has been influenced significantly by, and benefited greatly from, the many in-depth conversations that I have had with Dr. Sasha Outkin and Dr. Brandon Eames — for whose presence on my defense committee I am also grateful.

Thirdly, getting a PhD is challenging not merely from an intellectual perspective, but just as much so from an emotional perspective. I could not imagine completing my doctoral studies without my unwaveringly supportive community and family. At Georgia Tech, my

friends — Will Lassiter, Adrian Rivera, Ahmad Baubaid, Amanda Chu, and so many others — made going through coursework, studying for the comprehensive exams, and sharing ideas (both research and otherwise) as enjoyable as possible. I am immensely grateful for my community at Church of the Apostles, who welcomed me lovingly and have always encouraged me in my walk through life and with the Lord. There I have found genuine friends, who have celebrated with me my successes, supported me in my failures, and constantly remind me that there is more to living life abundantly than successes in my work.

Lastly, I owe so much of who I am today to the people who know and understand me best, and to whom I feel these words of gratitude are substantially insufficient. To my loving girlfriend, Taylor Kate Eubanks, thank you for your constant encouragement and support, for the joy and levity that you have brought to my life, and for making me feel understood. To my sister, Emily, and my brother, Seth, I am so proud of you two and who you have become. Emily, thank you for the unique dose of both wisdom and realism that you always know to give me; Seth, for your fun-loving and joyful sense of humor that reminds me to not take life too seriously. To Mom and Dad, I feel I cannot adequately express how much I owe to you. You never ceased to encourage me to live boldly and confidently pursue my dreams; you sacrificed your time and money so that I could have all the opportunities in the world to grow and succeed; you provided a stable and nurturing home life in which to grow up; and you have always loved me and supported me unconditionally. For all of that and so much more, I can only simply say thank you.

"*For to Him, and through Him, and for Him, are all things. To Him be the glory forever. Amen!*" (Romans 11:36)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation consists of three distinct, although conceptually related, papers that are unified in their focus on data-driven, stochastic sequential decision-making environments, but differentiated in their respective applications. In Chapter 2, we discuss a special class of partially observable Markov decision processes (POMDPs) in which the sources of uncertainty can be naturally separated into a hierarchy of effects — controllable, completely observable effects and exogenous, partially observable effects. For this class of POMDPs, we provide conditions under which value and policy function structural properties are inherited from an analogous class of MDPs, and discuss specialized solution procedures.

In Chapter 3, we discuss an inventory control problem in which actions are time-lagged, and there are three explicit sources of demand uncertainty — the state of the macroeconomy, product-specific demand variability, and information quality. We prove that a base stock policy — defined with respect to pipeline inventory and a Bayesian belief distribution over states of the macroeconomy — is optimal, and demonstrate how to compute these base stock levels efficiently using support vector machines and Monte Carlo simulation. Further, we show how to use these results to determine how best to strategically allocate capital toward a better information infrastructure or a more agile supply chain.

Finally, in Chapter 4, we consider how to generate trust in so-called development processes, such as supply chains, certain artificial intelligence systems, and maintenance processes, in which there can be adversarial manipulation and we must hedge against the risk of misapprehension of attacker objectives and resources. We show how to model dynamic agent interaction using a partially-observable Markov game (POMG) framework, and present a heuristic solution procedure, based on self-training concepts, for determining a robust defender policy.

CHAPTER 1
INTRODUCTION

This dissertation is unified in its focus on partially observable sequential decision-making environments under uncertainty. That is, this research is concerned with real-world environments in which a decision-maker (DM) is tasked with making decisions sequentially over time in such a way so as to optimize over some objective criterion, *i.e.* minimize the costs or maximize the rewards, such as revenue or profit, accrued over the decision horizon. Moreover, we consider decision-making environments that involve both *uncertainty* in their dynamics — that each decision DM induces the state of the system to change according to a random draw from a probability distribution — and *partial observability* in their information structure — that the DM must make decisions on the basis of *data* that inform, but do not completely reveal, the true state of certain aspects of the system. Since decisions are made on the basis of data, rather than direct observation (of at least some aspects) of the state of the system, these decision-making environments necessitate that the DM specify a way to *learn*, or process, data over time into decisions, so as to *optimize* with respect to their objective criterion. Solutions for the DM in such environments are, therefore, *policies* — functions that specify at each decision epoch what action should be taken, given the data available.

Decision-making environments such as these are manifest in much of human (and computer-based) experience. A company must make operational decisions on the basis of data about the micro- and macro-economic environment, so as to reach business objectives. A self-driving car must make decisions on the basis of sensor readings (data) of the state of the road, in order to reach its destination safely and efficiently. We all make decisions in our lives on the basis of what we know and observe, so as to increase our health, our happiness, our pleasure, *etc*. The wide applicability of methods for modeling

such decision-making environments broadly motivates our study in this dissertation.

This body of this dissertation is composed of three distinct papers that consider different partially observable sequential decision-making environments under uncertainty.

In Chapter 2, we consider a class of sequential decision-making problems under uncertainty in which there is a natural hierarchy of effects: micro-level forces that the decision-maker (DM) can control, and macro-level forces that the DM cannot. These problems have a completely observed state process subject to control of the DM and a partially observed modulation process exogenous to DM control that can affect the dynamics of the state process. We model this broad class of problems as a specially structured partially observed Markov decision process (POMDP) and call it the *modulated* POMDP or *M-POMDP*. The M-POMDP has many application areas, from inventory control to healthcare systems to dynamic pricing. By separating the belief update from actions taken, we show that the M-POMDP inherits value function and optimal policy function structure from its completely observed MDP analog. Further, we show that the M-POMDP allows for specialized approximate solution procedures based on solution procedures for the MDP.

In Chapter 3, we consider an inventory control problem that explicitly incorporates three different types of uncertainty — the state of the macroeconomy, product-specific demand variability, and information quality — in which ordering decisions are time-lagged and made on the basis of historical demand and noise-corrupted observational data that are Markov-modulated. We demonstrate how to efficiently compute an optimal ordering policy using grid-based approximation methods, simulated trajectories of future observational uncertainties, and successive support vector machines. We then consider how our model might address the following demand management question: should capital be allocated towards (1) a better information infrastructure (data, forecasts, quantitative talent, *etc.*), or (2) a more agile product architecture and supply chain design in order to more quickly respond to changing demand? We show that better information quality and agility improves system performance under optimal policies, and demonstrate how to numerically quantify these

effects using Monte Carlo simulation.

In Chapter 4, we consider how to increase trust in development processes in which there is risk for adversarial manipulation and the adversary's objectives, resources, and level of rationality are either ill-specified, imprecisely specified, or unknown. In such problems, we must hedge against the risk of misapprehension of attacker objectives, resources, and rationality, which is further complicated in the absence of adversarial training data. We show how to model dynamic agent interaction, on the basis of partially observed or noise corrupted data, using a partially-observable Markov game (POMG) framework. We then propose a three-fold heuristic solution procedure that (1) uses the POMG to generate potential adversarial policies, (2) explicitly incorporates these adversarial policies in the construction of a robust defender policy by solving a robust dynamic program, and (3) employs a probability matching heuristic in partially observable environments.

Finally, in Chapter 5, we discuss future research directions that naturally emanate from this work.

CHAPTER 2
SEQUENTIAL DECISION-MAKING AFFECTED BY PARTIALLY OBSERVED EXOGENOUS FORCES

2.1 Introduction

2.1.1 Motivation

Sequential decision-making environments often involve multiple tiers of effects — (1) the micro-level forces that the decision-maker (DM) can influence or control through actions, and (2) the macro-level forces that the DM cannot. This uncontrollable, macro-level tier of effects includes forces such as the macro-economy, the stock market, or natural phenoma, (*e.g.* weather, water currents, airstreams, *etc.*). These tiers of effects often also correspond to *degrees of observability*, leading to an information asymmetry. To reflect this environment, we present a specially structured partially observed Markov decision process (POMDP) having a completely observed state process (representing micro-level effects) and a partially observed modulation process (representing macro-level effects), where the modulation process affects the decision-making environment but is not affected by actions of the DM. We call this specially structured POMDP the *modulated POMDP*, or *M-POMDP*.

We remark that an important feature of the general POMDP is that it allows for the study of the interplay between information and control. Although the M-POMDP separates the belief update from the action and thus does not consider this interplay, the M-POMDP models a broad, far from restrictive, class of real-world problem settings where the DM is affected by forces that the DM cannot control but must take into account. Reality is such that: airlines must consider the weather in planning routes; investors must consider macro-economic conditions when making investment decisions; urgent, personalized healthcare

therapy manufacturers must consider the patient's health when making production decisions; fish must consider currents in swimming across a river. Our separation assumption is analogous to the fundamentally important principle of separation of estimation and control in optimal stochastic control, where an optimal observer for the state of the system does not depend on the choice of control ([6], [55]). The objective of this paper is to study the implications of this class of POMDPs.

Decision-making environments of this sort arise in numerous applications in the literature, without a unifying analytical framework, which justifies and inspires much of this research. These applications include inventory control, healthcare, and dynamic pricing.

Inventory Control. [31] considered a completely observed inventory control problem with reorder cost $K \geq 0$, augmented with a partially observed exogenous modulation process, assumed the demand process was dependent on the modulation process, and assumed the modulation process was observed by both the demand process and a so-called "additional observation data" (AOD) process. For example, the modulation process can model the underlying state of the economy; the AOD process can model various macro-economic indicators, *e.g.*, the number of housing starts, consumer spending, *etc*. This data-driven, modulated demand model generalizes models considered in the Markov-modulated demand and Bayesian updating literatures.

For this specially structured POMDP, (an M-POMDP, as we will show in Section 2.6) [31] showed that a generalized attainability assumption implied the existence of an optimal myopic base-stock policy if $K = 0$ and the existence of an optimal (s, S) policy if $K > 0$. [31] showed that (1) when $K = 0$, the value of the optimal base-stock level is constant within regions of the belief space and that these regions can be described by a finite set of linear inequalities, and (2) when $K > 0$, the values of s and S and upper and lower bounds on these values are constant within regions of the belief space and that these regions can be described by a finite set of linear inequalities. Further, under certain conditions, the base-stock levels are shown to be monotone with respect to a partial order on the belief

space.

[54] present a similar inventory control problem in which inventory levels are completely observed and the demand process is nonstationary and partially observed. In their formulation, replenishment time may be instantaneous or a fixed number of decision epochs. The probability distribution of demand is determined by the state of a Markov chain (which corresponds to a modulation process in our formulation). They derive that a base-stock policy, in which base-stock levels are parametrized by the belief distribution over the possible demand distribution states, is optimal and analyze the performance of various sub-optimal heuristics. As in [31], this model can also be cast as a M-POMDP.

Healthcare. [41] considered the problem of when patients with end-stage liver disease should accept or reject liver transplant offers. They model this problem as a completely observed Markov decision process (MDP), including two state processes, the liver transplant offer and the patient's rank on the transplant list, that are exogenous to control. The quality of the liver offered is dependent upon the patient's rank state. At each decision epoch, the patient may choose to accept or reject an offered liver, so they control (to some extent) their health state, which is dependent upon these exogenous liver offer and rank state processes. The authors prove various structural properties of the value function and optimal policy function, such as the monotonicity of the value function with respect to an order on the health states and the existence of an optimal control limit policy with respect to the liver offer quality. [42] is an extension of this model to the case in which the rank state is partially observed and, under the same rank state conditioning assumptions as in [41], is a M-POMDP. The partially observed model in [42] is shown to inherit structural properties from the completely observed model in [41], and features monotone structure of the value function and optimal policy function with respect to the belief space.

Dynamic Pricing. [2] introduced a dynamic pricing problem for highly seasonal products, in which demand is dependent upon pricing decisions and the underlying partially observed "core states" of their model, which are exogenous to control and (as in [31]) are

used to model product seasonality and demand correlations over time. Their initial formulation of the model is a POMDP, which they note is intractable to solve due to the intricate dependence of both the tasks of optimizing the objective function and learning the core states on the pricing decision. However, they propose a heuristic solution procedure in which they augment their information process with a hypothetical observation process in order to develop an approximation of their POMDP in which the task of learning the underlying core states is independent of action. This approximate model is a M-POMDP, and they observe that this approximate model features a passive learning environment, in which the task of learning the underlying core states of the model is independent of the actions taken to optimize the objective. This makes the approximate model much more tractable than the original POMDP model.

In this paper, we generalize the models presented in [31], [54], [41], [42], and [2] to a problem having a completely observed state process, a partially observed modulation process, an observation process, and an action process, where (i) the value of the state process at the next decision epoch depends on the current action selected, the current state process value, and the realization of the observation process at the next decision epoch, and (ii) the realizations of the observation and modulation processes at the next decision epoch are dependent on only the current value of the modulation process. Thus, the state process is affected by the modulation process only through the observation process, the observation process only observes the modulation process, and the modulation process is not affected by actions selected by the decision-maker.

The objective of this paper is to provide a unifying analytical framework for this broad class of models, by which we recover and generalize the salient features and structural properties found in [31], [54], [41], [42], and [2], and determine the implications in developing efficient solution procedures. To this effect:

(1) We show that the M-POMDP inherits value function and optimal policy function structure with respect to the completely observed state process from a set of closely

related (completely observed) MDPs.

(2) We show that the M-POMDP inherits value function structure with respect to the belief vector from the general POMDP, and discuss the value of information in M-POMDPs.

(3) We determine solution procedures that are based on the special structure of the M-POMDP.

2.1.2 Literature Review

The research towards (1) is inspired by [35] and [46], and by the commonalities of the structural results in [31], [54], [41], and [42]. [35] considered a notion of structure (which we adopt) as a restricted subspace of a function space in which every function in the subspace possesses some property of interest, and presented sufficient conditions by which a dynamic program has a value function and/or optimal policy function that are structured in this sense. We observe that structure has been useful for improved implementation and, as noted by [46], in developing a qualitative understanding of the model and characterizing how the results will vary with changes in model parameters. For example, the optimality of a base-stock policy for a large class of inventory control models is easy to implement and has significant impact computationally. Further, [46] showed that for a MDP, if the reward function satisfies a property $\mathscr{P}$ and the transition probabilities satisfy a stochastic version of property $\mathscr{P}$, then the value function satisfies property $\mathscr{P}$, where structural properties that satisfy property $\mathscr{P}$ include monotonicity, convexity, supermodularity, combinations of these, and other properties of interest. We remark that, whereas [46] only considers value function structure, we consider optimal policy structure as well. Additionally, we investigate structure with respect to the belief function, the value of information, and solution procedures that utilize special properties of the M-POMDP that are beyond the scope of the results in [46].

Further, we note that the results in [31], [54], [41], and [42] are such that the value function and optimal policy function are structured with respect to the completely observed state variable, and parametrized by the belief distribution over the partially observed state variable. In [31], the value function is convex with respect to the inventory on hand and has an optimal base-stock policy, in which the value function and base-stock level are parametrized by the belief distribution over the state of the broader economy. In [42], the value function is monotone with respect to the health of the patient and has an optimal control-limit policy, in which the value function and control-limit are parametrized by the belief distribution over the rank on the liver transplant list. The base-stock and control-limit policy structures are inherited from simpler models that are agnostic to the broader state of the economy and the rank on the liver transplant list, respectively.

The research towards (2) is inspired by various structural and monotonicity results in the POMDP literature ([45], [48], [29], [1], [58], [59], [61], [60]). [31] shows that the optimal base-stock levels are monotone in the belief vector under a first-order stochastic dominance assumption on demand. [42] reports similar monotonicity results. Further, we investigate the value of information accrued via observations. For POMDPs, more accurate observations of the underlying state process will never degrade optimal systems performance but may degrade sub-optimal systems performance [62]. We recover similar results for the M-POMDP in this paper, which differ from the general POMDP results in that we allow observations to impact state dynamics and cost.

Finally, the research towards (3) is motivated by the well-known problem with POMDPs that the belief space is uncountably infinite, leading to computational complications. Various solution approaches from exact methods ([45], [48], [25]), to fixed grid approximations ([30], [22]), to simulation-based approximations ([34], [49]) have been proposed. We present specialized approximate solution procedures for the M-POMDP:

- An *a priori* simulation-based grid approximation method that utilizes the structure of M-POMDP dynamics to separate the learning and optimizing tasks, a salient feature

of the solution approach in [2].

- A *real-time* heuristic procedure that exploits the relationship between the M-POMDP and its MDP analog.

- Approximation procedures based on information relaxation of the modulation state, as in [7].

We remark that the modulation and observation processes in our model may be equivalently viewed as a hidden Markov process (HMP), also referred to as a hidden Markov model (HMM). Thus, we may view our specialized class of POMDPs as MDPs with a state process that is weakly coupled to a HMP through observations. There is a thorough literature pertaining to the analytical and asymptotic properties, parameter estimation, and applications of HMPs. The literature of HMPs is useful for estimating the M-POMDP model parameters in real-world applications. For an introduction to this literature, we refer the reader to the survey [16].

2.1.3 Outline

The paper is organized as follows. In Section 2.2 we present a POMDP with a completely observed state process, a partially observed modulation process, an observation process, and an action process. We present preliminary results for this POMDP in Section 2.3, and begin by applying known results to the POMDP defined in Section 2.2 that lead to the development of the optimality equation. We then make key conditioning assumptions, which leads to the definition of the M-POMDP. These assumptions lead to a reformulation of the optimality equation and the definition of a set of completely observed MDPs that we call the MDP analog to the M-POMDP. We then extend three structural conditions due to [35] to the M-POMDP, followed by an extension of the structural results of [46] to the M-POMDP.

We present our main structural results in Section 2.4. In this section, we first charac-

terize the *inheritance property* of M-POMDPs, giving sufficient conditions by which the M-POMDP inherits structure in its value function and optimal policy, for a large class of structural properties, from its MDP analog. We then address the structural properties of the value function with respect to the belief vector over the modulation space. We show that the M-POMDP inherits concavity of the value function from the general POMDP, and discuss the value of information in M-POMDPs. Additionally, we present analogous results when state dynamics are described by a difference equation, as is common for inventory problems, rather than by conditional probabilities.

Section 2.5 gives examples of how our results may be used to extend MDP structural results to the M-POMDP. We show in Section 2.5.1 that the M-POMDP inherits monotone optimal policy structure from its MDP analog. In Section 2.5.2, optimal myopic policy structure is shown to be inherited by the M-POMDP by determining that separability is a structure that satisfies results in Section 2.3. In the appendix, we additionally consider $L^\natural$-convexity and multi-modularity as C3 properties, which are structures present in certain inventory problem settings ([28], [64]).

Section 2.6 demonstrates an example of how the M-POMDP (albeit not under this name) has been used in the literature. We consider the inventory control problem of [31] in Section 2.6.1 and show that M-POMDPs inherit the optimality of a base-stock policy for the case where there is no reordering cost and backlogging is allowed.

Section 2.7 presents two solution procedures for the M-POMDP that do not extend to the more general POMDP. In Section 2.7.1, we present an *a priori* simulation-based approximation method based on the dynamics of the modulation process to transform the M-POMDP into an MDP. In Section 2.7.2, we present a heuristic approach that transforms the M-POMDP into an MDP at each decision epoch, each of which is more tractable than the MDP derived in Section 2.7.1. We then discuss solution procedures using an information relaxation lower bound in Section 2.7.3 and the tradeoffs between our approaches in Section 2.7.4. Conclusions and directions for future research are presented in Section 2.8.

2.2 Problem Formulation

Consider a POMDP that has an infinite horizon and discrete decision epochs $t = 0, 1, \ldots$, and involves a completely observed state process $\{s_t, t \geq 0\}$ existing in a space S, a partially observed modulation process $\{\mu_t, t \geq 0\}$ in a space M, an observation process $\{z_t, t \geq 1\}$ in a space Z, and an action process $\{a_t, t \geq 0\}$ in a space $A = \bigcup_{s \in S} A(s)$, where $a_t \in A(s_t), \forall t$. Assume that S, M, Z, A are discrete spaces and that these processes are linked by the conditional probability $P[z_{t+1}, s_{t+1}, \mu_{t+1} | s_t, \mu_t, a_t]$. It will be convenient for notational purposes to let $P[z_{t+1}, s_{t+1}, \mu_{t+1} | s_t, \mu_t, a_t] = P[z', s', \mu' | s, \mu, a]$.

We assume that $c : S \times Z \times A \mapsto \mathbb{R}$ is the bounded single period cost function, where $c(s_t, z_{t+1}, a_t) = c(s, z', a)$ is the cost accrued during period $[t, t+1)$. We further assume that the action at epoch t can be selected on the basis of the information received up to t, $\mathscr{I}_t = \{s_t, s_{t-1}, \ldots, s_0, z_t, z_{t-1}, \ldots, z_1, a_{t-1}, a_{t-2}, \ldots, a_0, x_0\}$, where $x_0 = \{x_0(\mu), \mu \in M\}$ and $x_0(\mu) = P[\mu_0 = \mu]$ for all $\mu \in M$. A function mapping the set of all $\mathscr{I}_t$ into the set of all actions for all t is a feasible policy. The problem criterion is the expected total discounted cost over the infinite horizon, where we assume β, $0 \leq \beta < 1$ is the discount factor. The problem is to determine a feasible policy that minimizes the criterion with respect to all feasible policies.

2.3 Preliminary Results

Results in [45] and [48] imply that $\{(s_t, x_t), t \geq 0\}$ is a sufficient statistic for this problem, where $x_t = \{x_t(\mu), \mu \in M\}$ and $x_t(\mu) = P[\mu_t = \mu | \mathscr{I}_t]$. We call x_t the Bayesian belief

function at epoch t and $\{x_t, t \geq 0\}$ the belief function process. Let

$$\begin{aligned}
\theta(z', s'|s, x, a) &= \sum_{\mu'} \sum_{\mu} x(\mu) P[z', s', \mu'|s, \mu, a] \\
\lambda(\mu'|z', s', s, x, a) &= \frac{\sum_{\mu} x(\mu) P[z', s', \mu'|s', \mu, a]}{\theta(z', s'|s, x, a)}, \quad \theta(z', s'|s, x, a) \neq 0 \\
\lambda(z', s', s, x, a) &= \{\lambda(\mu'|z', s', s, x, a), \mu' \in M\}.
\end{aligned}$$

We can think of $\lambda(z', s', s, x, a)$ as the posterior belief function x_{t+1}, given $x_t = x, a_t = a, s_t = s, s_{t+1} = s'$, and $z_{t+1} = z'$. Similarly, $\theta(z', s'|s, x, a)$ is the probability that $z_{t+1} = z'$ and $s_{t+1} = s'$, given that $s_t = s$, $x_t = x$, and $a_t = a$. Let V be the Banach space of bounded value functions which map $S \times X$ into $\mathbb{R}$ endowed with the sup-norm, and let $H : V \mapsto V$ be defined as

$$Hv(s, x) = \min_{a \in A(s)} \left\{ \mathbb{E}\left[c(s, z', a)|x\right] + \beta \sum_{z', s'} \theta(z', s'|s, x, a) v\big(s', \lambda(z', s', s, x, a)\big) \right\}, \quad (2.1)$$

where $\mathbb{E}\left[c(s, z', a)|x\right] = \sum_{z', s'} \theta(z', s'|s, x, a) c(s, z', a)$. The optimality equation is $v = Hv$. Results from [37] guarantee, by the contraction property of H, the existence of a unique value function, v^*, such that $v^* = Hv^*$, and that this fixed point is the expected total discounted cost accrued by an optimal policy. Further, we can restrict search for an optimal policy to t-invariant functions that select a_t on the basis of s_t and x_t. Let Π to be the space of such t-invariant functions from $S \times X$ to A. The function, $\pi \in \Pi$ such that $\pi(s_t, x_t) = a_t$ causing the minimum in equation (2.1) to be attained is an optimal policy. The expected total discounted cost accrued by this optimal policy can be attained by recursive application of H, so that $\lim_{n \to \infty} \|v^* - v_n\| = 0$, where $v_{n+1} = Hv_n$ for all n, given v_0 is any function in V, and $\|\cdot\|$ is the sup-norm.

2.3.1 Key Conditioning Assumptions.

By the definition of conditional probability,

$$P[z', s', \mu' | s, \mu, a] = P[s' | z', \mu', s, \mu, a] P[z', \mu' | s, \mu, a].$$

We assume that

$$P[s' | z', \mu', s, \mu, a] = P[s' | z', s, a] \tag{2.2}$$

$$P[z', \mu' | s, \mu, a] = P[z', \mu' | \mu]. \tag{2.3}$$

We call the POMDP presented in Section 2.2 with these key conditioning assumptions the *modulated POMDP*, or the M-POMDP.

We remark that the standard POMDP definition in the literature ([45], [48]) assumes three processes, the partially observed state process, the observation process, and the action process, all of which are linked by the given probability $P[z', s' | s, a]$. This standard definition assumes $P[z', s' | s, a] = P[z' | s', s, a] P[s' | s, a]$, where $P[z' | s', s, a]$ describes the relationship between the state, observation, and action processes and $P[s' | s, a]$ describes the controlled dynamics of the state process. We note that the conditioning for the POMDP considered in this paper, $P[z', s', \mu' | s, \mu, a] = P[s' | z', s, a] P[z', \mu' | \mu]$, assumes that s' is dependent on z', rather than vice versa.

Thus, for the M-POMDP we assume that the state process is affected by the modulation process only through the observation process, the observation process only observes the modulation process, and the modulation process is exogenous to control. Under these

assumptions, we can rewrite θ,

$$\begin{aligned}\theta(z', s'|s, x, a) &= \sum_{\mu} x(\mu) \sum_{\mu'} p(s'|z', s, a) P[z', \mu'|\mu] \\ &= p(s'|z', s, a) \sum_{\mu,\mu'} x(\mu) P[z', \mu'|\mu] \\ &= p(s'|z', s, a)\sigma(z'|x),\end{aligned}$$

where we let $p(s'|z', s, a) = P[s'|z', s, a]$, and $\sigma(z'|x) = \sum_{\mu,\mu'} x(\mu)P[z', \mu'|\mu]$. We can then rewrite λ, by plugging in for θ and assuming $\theta(z', s'|s, x, a) \neq 0$, as follows:

$$\begin{aligned}\lambda(\mu'|z', s', s, x, a) &= \frac{\sum_{\mu} x(\mu) P[z', s', \mu'|s, \mu, a]}{\theta(z', s'|s, x, a)} \\ &= \frac{\sum_{\mu} x(\mu) P[z', s', \mu'|s, \mu, a]}{p(s'|s, z', a)\sigma(z'|x)} \\ &= \frac{\sum_{\mu} x(\mu) P[z', \mu'|\mu]}{\sigma(z'|x)}.\end{aligned}$$

Thus, $\lambda(\mu'|z', s', s, x, a)$ is independent of s', s, a, and we denote $\lambda(\mu'|z', s', s, x, a) = \lambda(\mu'|z', x)$ for all $\mu' \in M$ and $\lambda(z', x) = \{\lambda(\mu'|z', x), \mu' \in M\}$.

Note $\mathbb{E}[c(s, z', a)|x] = \sum_{z',\mu'} \sum_{\mu} P[z', \mu'|\mu]x(\mu)c(s, z', a) = \sum_{z'} \sigma(z'|x)c(s, z', a)$, and let

$$h_{z'}(s, a, \bar{v}) = c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a)\bar{v}(s').$$

We then reformulate the operator H as follows:

$$Hv(s, x) = \min_{a \in A(s)} \left\{ \sum_{z'} \sigma(z'|x) h_{z'}\big(s, a, v(\cdot, \lambda(z', x))\big) \right\}.$$

We can now define the completely observed MDP analog to the M-POMDP. Let $MDP_{z'}$ have single period cost function $c(s, z', a)$, transition structure $\{p(s'|z', s, a)\}$, and operator

$$\bar{H}_{z'}\bar{v}(s) = \min_{a \in A(s)} h_{z'}(s, a, \bar{v}). \tag{2.4}$$

We call the collection $\{MDP_{z'} : z' \in Z\}$ the completely observed MDP analog of the M-POMDP.

2.3.2 The Porteus Results Extended

Let V_x denote the halfspace of V induced by affixing $x \in X$ (*i.e.* $V_x = \{f(\cdot, x) : f \in V\}$, $\forall x \in X$) and Π_x denote the halfspace of Π induced by affixing $x \in X$. Suppose $\tilde{V}$ is a space of structured value functions $S \mapsto \mathbb{R}$, and $\tilde{\Pi}$ is a space of structured Markovian deterministic policy functions $S \mapsto A$.

We now present the three structural conditions found in [35] extended to the M-POMDP setting:

P(a) Structured space of functions contains its limit points

$\tilde{V}$ is a closed subset of $V_x, \forall x \in X$.

P(b) Structured Value Preservation

$v(\cdot, x) \in \tilde{V}, \forall x \in X \Rightarrow Hv(\cdot, x) \in \tilde{V}, \forall x \in X$.

P(c) Structured Policy Attainment

$$v(\cdot, x) \in \tilde{V}, \forall x \in X \;\Rightarrow \exists \pi(\cdot, x) \in \tilde{\Pi}, \forall x \in X \text{ s.t.}$$
$$Hv(\cdot, x) = \sum_{z'} \sigma(z'|x) h_{z'}\big(\cdot, \pi(\cdot, x), v(\cdot, \lambda(z', x))\big), \forall x \in X.$$

We refer to P(a), P(b), and P(c) as the *extended Porteus conditions*. Condition P(a) insures that the limit point of a sequence of value functions obtained by the value iteration algorithm will be in the space of structured value functions, condition P(b) insures that the structure of the value function is preserved when applying the dynamic programming operator H, and condition P(c) insures that for all structured value functions on S, it suffices to search the space of structured policies (smaller than the space of all policies) for a v-improving policy.

We present a proposition in which we establish that P(a), P(b), and P(c) are sufficient conditions to guarantee that the value function and an optimal policy function are structured on S. Subsequent results pertaining to structure on S demonstrate sufficient conditions for P(a), P(b), and P(c) to hold, by investigating the M-POMDP model primitives and the relationship to the MDP analog.

Proposition 1. *Assume the extended Porteus conditions hold. Then there exists a* $\pi^*(\cdot, x) \in \tilde{\Pi}$ *and a* $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$ *such that*

$$v^*(s,x) = Hv^*(s,x) = \sum_{z'} \sigma(z'|x) h_{z'}\big(s, \pi^*(s,x), v^*(\cdot, \lambda(z',x))\big)$$

for all $(s,x) \in S \times X$.

Proof of the above result is a straightforward extension of Theorem 6.11.1 in [37]. We remark that the structured optimal value function and the structured optimal policy are both modulated by the belief process $\{x_t, t > 0\}$. The following corollary establishes that it is sufficient for only P(a) and P(b) to hold to establish structure of the value function on S, absent structure in the policy.

Corollary 1. *If only P(a) and P(b) hold, then* $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$.

2.3.3 Smith & McCardle Results Extended

We now seek properties that guarantee value function structure on S. Our approach, in this subsection, is to extend properties due to [46] to the M-POMDP. Towards this aim, we present four definitions due to [46] — the *closed convex cone property* (C3), the *single point property*, the *joint extension*, and the *stochastic dominance analog* — and three related results — Propositions 2, 3, and 4 — that work towards the structural results in Section 2.4. We begin with the definition of a C3 property.

Definition 1. *(C3 property)* $\mathscr{P}$ *is a closed convex cone property (C3) if and only if the set*

of all real-valued functions on S satisfying $\mathscr{P}$ forms a closed convex cone in the topology of pointwise convergence.

We assume for the remainder of this section that $\tilde{V}$ is a space of structured value functions possessing a C3 property, $\mathscr{P}$. Proposition 1 in [46] gives us an equivalent definition of C3 property in terms of an inequality "test of satisfaction". A real-valued function f on S satisfies a C3 property if and only if there exists a finite set of points $\{s_j, j \in J_k\}$, $\{s_i, i \in I_k\}$ and positive weights $\{\gamma_j, j \in J_k\}$ and $\{\gamma_i, i \in I_k\}$ such that

$$\sum_{j \in J_k} \gamma_j f(s_j) \leq \sum_{i \in I_k} \gamma_i f(s_i), \quad \forall k \in K$$

where K is an index set.

For Proposition 2 below, we will need to consider a special sub-class of C3 properties, defined in terms of the "test of satisfaction" definition of C3 properties.

Definition 2. *(Single Point Property) A C3 property $\mathscr{P}$ is considered a single-point property if and only if any function $g : S \mapsto \mathbb{R}$ with $\mathscr{P}$ has an inequality test of satisfaction with the following form*

$$g(s_k) \leq \sum_{i \in I_k} \gamma_i g(s_i), \quad \forall k \in K$$

where $\{\gamma_i, i \in I_k\}$ is a finite set of positive weights, and K is an index set.

Many of the structural properties that we care about are in fact single point properties: isotonicity, antitonicity, convexity, subadditivity. The notion of single-point properties encapsulates the intuition that if the function, $f : S \times A \mapsto \mathbb{R}$ to be minimized satisfies a structural property on S for every $a \in A$, then the function, optimized over A should retain the property on S. As [46] show, this intuition does not hold in general, but does hold for this special class of single-point C3 properties.

Our next result is the first step towards extending MDP results to the M-POMDP. The idea is if $h_{z'}(s, a, v)$ has a single-point C3 property on S, for all $a \in A$ and $z' \in Z$, then the optimal value function v^* will have structure on S, for each belief state $x \in X$.

Proposition 2. *Assume for all $\bar{v} \in \tilde{V}$, $-\bar{v}$ possesses a special single-point C3 property, $\mathscr{P}$. If $v(\cdot, x) \in \tilde{V}$ for all $x \in X$ implies $h_{z'}(\cdot, a, v(\cdot, x)) \in \tilde{V}$, for all $a \in A, x \in X$, then $v^*(\cdot, x) \in \tilde{V}$ for all $x \in X$.*

Our next result provides another set of sufficient conditions on $h_{z'}$ for which P(b) holds. In fact, these sufficient conditions contain the prior conditions. Towards this effort, we introduce the notion of a *joint extension* of a C3 property.

Definition 3. *(Joint Extension) Given a C3 property $\mathscr{P}$ on S, a function $f : S \times A \mapsto \mathbb{R}$ satisfies a joint extension of $\mathscr{P}$ on $S \times A$, call it $\mathscr{P}^*$, if and only if for any $k \in K$, actions $\{a_j, j \in J_k\}$, $\exists\{a_i, i \in I_k\}$ such that*

$$\sum_{j \in J_k} \gamma_j f(s_j, a_j) \leq \sum_{i \in I_k} \gamma_i f(s_i, a_i)$$

where $\{\gamma_j, j \in J_k\}$, $\{\gamma_i, i \in I_k\}$ are finite sets of positive weights associated with the test of satisfaction for $\mathscr{P}$.

This next proposition makes use of Proposition 4 in [46] and states that if for $MDP_{z'}$, when $\bar{v}$ satisfies a C3 property and $h_{z'}(\cdot, \cdot, \bar{v})$ satisfies a joint extension of that C3 property for all $z' \in Z$, then v^* is structured on S.

Proposition 3. *If $\bar{v} \in \tilde{V}$ implies $h_{z'}(s, a, \bar{v})$ satisfies a joint extension of $\mathscr{P}$ in (s, a) on $S \times A$, $\mathscr{P}^*$, for all $z' \in Z$, then $v^*(\cdot, x) \in \tilde{V}$ for all $x \in X$.*

In our next result, we present conditions on the model primitives c, p, and v that guarantee P(b) holds, thereby guaranteeing that v^* is structured on S by Corollary 1. Towards this aim, we introduce the notion of stochastic dominance relations, which will define stochastic analogs for the C3 properties and their joint extensions that we have introduced thus far. We frame the definition explicitly in terms of the random variable parameters of interest in this paper. A more thorough and general treatment can be found in [46].

Let $s_{t+1}(s, z', a)$ be the random variable for the state at time $t+1$, conditioned on $s_t = s$, $z_{t+1} = z'$, and $a_t = a$. The transition probability function $p(\cdot|z', s, a)$ represents a probability measure on S, for all $s \in S, z' \in Z, a \in A$.

Definition 4. *(Stochastic dominance analog) We say that $s_{t+1}(z', s, a)$ stochastically dominates $s_{t+1}(z', s', a')$ on $\tilde{V}$, if*

$$\mathbb{E}v\left(s_{t+1}(z', s, a)\right) \geq \mathbb{E}v\left(s_{t+1}(z', s', a')\right), \quad \forall v \in \tilde{V}.$$

Additionally, [46] present an equivalent representation of these dominance relations with respect to the probability measures p, using a binary relation $\precsim_{\tilde{V}}$, such that $s_{t+1}(z', s, a)$ stochastically dominates $s_{t+1}(z'', s', a')$ on $\tilde{V}$ if $p(\cdot|z', s', a') \precsim_{\tilde{V}} p(\cdot|z', s, a)$. As in [46], we say that $s_{t+1}(z', s', a')$ satisfies the stochastic analog of $\mathscr{P}$ (a property of functions on $S \times A$) on $\tilde{V}$, call it $\mathscr{P}_{\tilde{V}}$, if the inequality test of satisfaction for $\mathscr{P}$ is satisfied with respect to $\precsim_{\tilde{V}}$.

This brings us to our next result, which is an extension of Proposition 5 in [46], and states that if the value, cost, and probability transition functions satisfy a C3 property, its joint extension, and its stochastic joint extension, respectively, for each observation z' and belief distribution x, then v^* is structured on S.

Proposition 4. *Suppose $\tilde{V}, \tilde{C}, \tilde{P}$ are spaces of structured value, cost, and probability transition functions for which $\bar{v} \in \tilde{V}$ has a C3 property, $\mathscr{P}$, $\bar{c} \in \tilde{C}$ has a joint extension of $\mathscr{P}$ on $S \times A$, and $\bar{p} \in \tilde{P}$ has a stochastic joint extension of $\mathscr{P}$ (a property of functions on $S \times A$) on $\tilde{V}$, call it $\mathscr{P}_{\tilde{V}}$. If the following conditions hold:*

(i) $c(\cdot, z', \cdot) \in \tilde{C}$ *for all* $z' \in Z$

(ii) $p(\cdot|z', \cdot, \cdot) \in \tilde{P}$ *for all* $z' \in Z$.

Then, $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$.

2.4 Main Structural Results

Propositions 2, 3, and 4 present sufficient conditions for guaranteeing value function structure on S by guaranteeing that P(b) holds under operator H. We now present our primary structural results, which formalize the *inheritance property* of M-POMDPs — that value function *and* optimal policy function structure of the MDP analog are inherited by the M-POMDP. Oftentimes in modeling efforts we make stylized and unrealistic simplifying assumptions for the sake of analytical tractability and gaining important qualitative intuition about a system (*e.g.* demand is *i.i.d.* across decision epochs, a firm operates independent of competitors). The thrust of the results in this section is that, for an important class of properties and models, we may analyze a simpler model and guarantee the structural properties hold for a more robust model. Thus, analytical tractability need not be traded for modeling realism. We use this inheritance property liberally in Sections 2.5 and 2.6 in order to give a flavor of how the results in this section may be used in extending general MDP results and in specific applications. In this section, we also demonstrate how many of the value function and optimal policy function structural results on X from the POMDP literature apply to the M-POMDP. We then, finally, relate the definition of dynamics via conditional probabilities to the functional description of dynamics, as is more natural in many application settings, such as inventory control.

2.4.1 Structure on S

We begin by stating the Porteus conditions for MDPs, and recapitulating, for ease of reference, the structural implications for the MDP analog.

P$_{z'}$(b) Structured Value Preservation

$$\tilde{v} \in \tilde{V} \Rightarrow \bar{H}_{z'}\tilde{v} \in \tilde{V}.$$

P$_{z'}$(c) Structured Policy Attainment

$$\tilde{v} \in \tilde{V} \Rightarrow \exists \tilde{\pi} \in \tilde{\Pi} \text{ s.t. } \bar{H}_{z'}\tilde{v} = h_{z'}(\cdot, \tilde{\pi}, \tilde{v}).$$

The following proposition is due to [35] and is stated in Theorem 6.11.1 in [37].

Proposition 5. *Suppose P(a), $P_{z'}$(b), and $P_{z'}$(c) hold. Then there exists a $\pi^*_{z'} \in \tilde{\Pi}$ and a $v^*_{z'} \in \tilde{V}$ such that $v^*_{z'}(s) = \bar{H}_{z'} v^*_{z'}(s) = h_{z'}\big(s, \pi^*_{z'}(s), v^*_{z'}\big)$, for all $s \in S$.*

Corollary 2. *Suppose P(a) and $P_{z'}$(b) hold. Then $v^*_{z'} \in \tilde{V}$.*

Suppose $\tilde{F}$ is a space of functions from $S \times A$ to $\mathbb{R}$ that satisfy a joint C3 property, $\mathscr{P}^*$. Further, let Δ be the space of feasible MDP analog policies from S to A (note that $\tilde{\Pi} \subseteq \Delta$). We present conditions by which the M-POMDP *inherits* this MDP analog structure:

B(a) $\tilde{v} \in \tilde{V} \Rightarrow h_{z'}(\cdot, \cdot, \tilde{v}) \in \tilde{F}$

B(b) $f \in \tilde{F} \Rightarrow \min_{\delta \in \Delta} f^\delta \in \tilde{V}$

B(c) $f \in \tilde{F} \Rightarrow \exists \tilde{\pi} \in \tilde{\Pi}$ s.t. $\min_{\delta \in \Delta} f^\delta = f^{\tilde{\pi}}$,

where $f^\delta(s) = f(s, \delta(s))$ for all $s \in S$, and the minimum with respect to $\delta \in \Delta$ is taken pointwise, *i.e.* $\left[\min_{\delta \in \Delta} f^\delta\right](s) = \min_{a \in A(s)} f(s, a)$ for all $s \in S$.

Condition B(a) guarantees that, for the MDP analog, the Bellman minimized function $h_{z'}$ is structured on $S \times A$. We recognize that this structure must be preserved under expectation in order for the M-POMDP Bellman minimized function to inherit this structure, which is guaranteed in that $\tilde{F}$ is a space of functions possessing a joint C3 property, $\mathscr{P}^*$. Condition B(b) insures that the minimization operation over feasible policies maps functions from $\tilde{F}$ into $\tilde{V}$. Finally, condition B(c) supposes we know, or can show, that minimizing functions of a certain structure on $S \times A$ yields a structured optimal policy. In fact, these conditions are quite mild, and hold for every one of the applications in Sections 2.5 and 2.6. There are various results in the literature in this vein, *e.g.* results pertaining to minimizing submodular functions on a lattice ([53]) and minimizing $L^\natural$-convex functions ([64]).

Note that B(a) and B(b) straightforwardly imply that $P_{z'}$(b) holds for all $z' \in Z$, and B(a) and B(c) imply that $P_{z'}$(c) holds for all $z' \in Z$. Thus, these are sufficient conditions for guaranteeing that the MDP analog is structured in its value function and an optimal policy

by Proposition 5. Our next proposition formalizes the *inheritance property* of M-POMDPs by demonstrating that these sufficient conditions for guaranteeing structure for the MDP analog are, in fact, also sufficient for guaranteeing the M-POMDP is structured on S in the same way. The proof follows by demonstrating that B(a), B(b), and B(c) are sufficient for guaranteeing that P(b) and P(c) hold, and then applying Proposition 1.

Proposition 6. *Suppose P(a), B(a), B(b), and B(c) hold. Then there exists a* $\pi^*(\cdot, x) \in \tilde{\Pi}$ *and a* $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$ *such that*

$$v^*(s,x) = Hv^*(s,x) = \sum_{z'} \sigma(z'|x) h_{z'}\big(s, \pi^*(s,x), v^*(\cdot, \lambda(z',x))\big)$$

for all $(s,x) \in S \times X$.

The following is a straightforward corollary that shows that P(a), B(a), and B(b) are sufficient for guaranteeing value function structure, absent policy structure.

Corollary 3. *Suppose P(a), B(a), and B(b) hold. Then* $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$.

Of course, if the model primitives $p = \{p(s|z',s,a)\}$ and $c = \{c(s',z',a)\}$ guarantee that B(a) and B(b) hold, then the M-POMDP is structured in its value function by the Corollary 3.

Corollary 4. *Suppose P(a) holds, and that* $p \in \tilde{P}$ *for all* $z' \in Z$ *and* $c \in \tilde{C}$ *for all* $z' \in Z$ *imply that B(a) and B(b) hold. Then* $v^*(\cdot, x) \in \tilde{V}$ *for all* $x \in X$.

2.4.2 Structure on X

In this subsection, we discuss some known structural properties related to POMDPs, as they pertain to the M-POMDP. The following proposition is due to [45] and [48], in which successive value approximations achieved by applying the Bellman operator, H, preserve piecewise linearity and concavity of v with respect to x. Concavity is preserved in the limit.

Proposition 7. *The value function* $v^*(s,\cdot)$ *is concave in* x *on* X*, for all* $s \in S$.

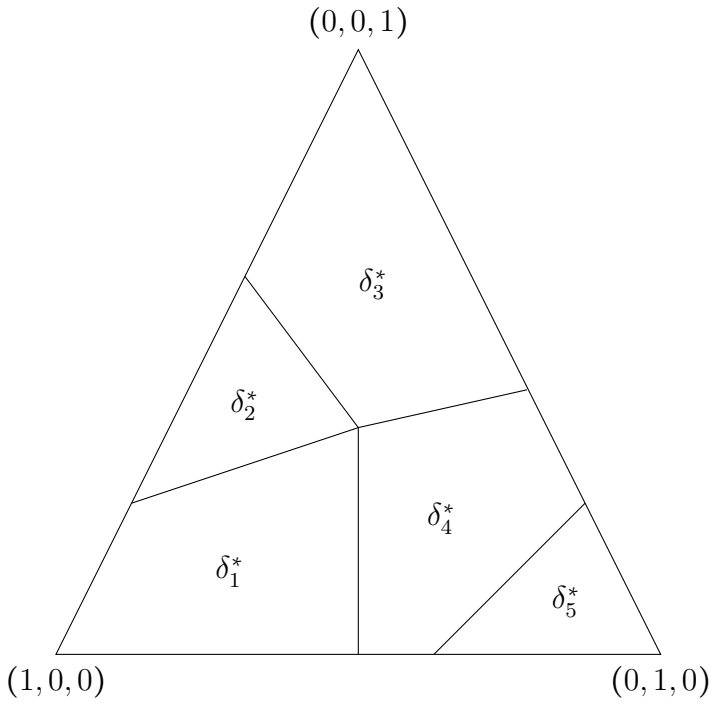


Figure 2.1: A graphical depiction of Corollary 5, with a 3-dimensional belief simplex X, and where $\pi^*(\cdot, x) = \delta_j^*$ for all x in partition region X_j.

If v^* can be shown to be piecewise linear in x on X as well (such as if the optimal policy is finitely transient, as in [48]), then we have a corollary result. For the standard POMDP model, the belief space X partitions into a finite number of convex, polyhedral regions that specify an optimal *control* or *action* to take. We note that for the M-POMDP, the belief space partitions into a finite number of convex, polyhedral regions that specify an optimal control or action *for each* $s \in S$. Thus, these non-overlapping regions in X specify a *partial policy*, *i.e.* functions from the state space S into the action space A. If Proposition 1 holds, then these regions specify *structured* partial policies.

Corollary 5. *Suppose v^* is piecewise linear in x on X. Then, there exists a partition of X into a finite number of convex, polyhedral regions $\{X_j, j = 1, \ldots, n\}$ such that there exists a set of functions from S into A, $\{\delta_j^*, j = 1, \ldots, n\}$, such that $\pi^*(\cdot, x) = \delta_j^*$ for all $x \in X_j$, $j = 1, \ldots, n$.*

The various monotonicity results for the standard POMDP relate to the M-POMDP, and

the following proposition is in the vein of results found in [29]. We will need the following notions of stochastic partial order (in increasing strength).

- *First-order stochastic dominance.* $x \geq_S$ if an only if $\sum_{i \geq q} x(i) \geq \sum_{i \geq q} x'(i)$.
- *Monotone likelihood ratio (MLR).* $x \geq_{LR} x'$ if and only if $x(i)x'(i') \geq x(i')x'(i)$, for all $i, i' \in M$.
- *Strong MLR.* For any two functions f, g from $Z \times M \to \mathbb{R}$, $f \geq_{TP} g$ if and only if $f(z \vee z', \mu \vee \mu')g(z \wedge z', \mu \wedge \mu') \geq f(z, \mu)g(z', \mu')$, where $z \vee z'$ indicates $\max\{z, z'\}$ and $z \wedge z'$ indicates $\min\{z, z'\}$.

We say that f is TP_2 if $f \geq_{TP} f$.

Proposition 8. *Suppose the following hold:*

(a) $c(s, z', a)$ *is nondecreasing in* z' *on* Z*, for all* $(s, a) \in S \times A$

(b) $p(s'|z', s, a)$ *is nondecreasing in* z' *on* Z*, for all* $(s', s, a) \in S \times S \times A$

(c) The $M \times M$ *matrix* $P[\cdot, z'|\cdot]$ *is* TP_2 *for all* $z' \in Z$

(d) $\sum_{\mu'} P[z', \mu'|\mu] \geq \sum_{\mu'} P[z', \mu'|\bar{\mu}]$ *for all* $z' \in Z$*,* $\mu \geq \bar{\mu} \in M$

(e) $P[z', \cdot|\cdot] \leq_{TP} P[\bar{z}, \cdot|\cdot]$ *for all* $z \geq \bar{z} \in Z$.

Then, $\bar{x} \leq_{LR} x$ *implies* $v^*(s, \bar{x}) \leq v^*(s, x)$ *for all* $s \in S$.

2.4.3 Value of Information

For the general POMDP, improved observation quality will not degrade performance if an optimal policy is applied, although this may not be true for sub-optimal policies ([45], [62]). The analogous result for the M-POMDP is not as straightforward due to the non-standard conditioning assumptions made in Section 2.3.1, as we now show.

Consider two M-POMDPs, the first of which has cost structure $\{\tilde{c}(s,\tilde{z},a)\}$, state transition structure $\{\tilde{p}(s'|\tilde{z},s,a)\}$, and observation structure $\{\tilde{q}(\tilde{z}|\mu',\mu)\}$, where $\tilde{q}(\tilde{z}|\mu',\mu) = \tilde{P}[\tilde{z}|\mu',\mu]$. The second of which has cost structure $\{c(s,z',a)\}$, state transition structure $\{p(s'|z',s,a)\}$, and observation structure $\{q(z'|\mu',\mu)\}$, where $q(z'|\mu',\mu) = P[z'|\mu',\mu]$. We assume both M-POMDPs share the same modulation dynamics $P[\mu'|\mu]$.

We now present a definition of improved observation quality, as presented in [45], [62], and elsewhere.

Definition 5. *For any two probability distributions $\tilde{q}(\tilde{z}|\mu',\mu), q(z'|\mu',\mu)$ over Z, we say that $\tilde{q}$ has improved observation quality over q if there exists a $Z \times Z$ stochastic matrix, ξ, such that $q(z'|\mu',\mu) = \sum_{\tilde{z}} \xi(z'|\tilde{z})\tilde{q}(\tilde{z}|\mu',\mu)$.*

We may view the stochastic matrix ξ as a Markov noisy channel, so that we consider an observation distribution $\tilde{q}$ improved relative to q if q is equivalent to receiving a signal from $\tilde{q}$ passed through a noisy channel. Let $\tilde{H}$ and H be the Bellman operators for the first and second problems, respectively. Assume $\tilde{v}_0$ and v_0 are given, $\tilde{v}_{n+1} = \tilde{H}\tilde{v}_n$ and $v_{n+1} = Hv_n$, and assume $\tilde{v}$ and v^* are the fixed points of $\tilde{H}$ and H, respectively. In the following proposition, we seek to use the concavity of the value function (Proposition 7) in order to demonstrate the relationship between improved observation quality and system performance under an optimal policy.

Proposition 9. *Suppose $\tilde{q}$ has improved observation quality over q, so that there exists a Markov noisy channel, ξ, such that $q(z'|\mu',\mu) = \sum_{\tilde{z}} \xi(z'|\tilde{z})\tilde{q}(\tilde{z}|\mu',\mu)$ for all $z' \in Z$, $\mu,\mu' \in M$. Assume:*

(a) $\tilde{c}(s,\tilde{z},a) = \sum_{z'} \xi(z'|\tilde{z})c(s,z',a)$, for all $(s,\tilde{z},a) \in S \times Z \times A$

(b) $\tilde{p}(s'|\tilde{z},s,a) = \sum_{z'} \xi(z'|\tilde{z})p(s'|z',s,a)$, for all $(s',s,\tilde{z},a) \in S \times S \times Z \times A$.

If $\tilde{v}_0 \leq v_0$, then $\tilde{v}_n \leq v_n$ for all n, and $\tilde{v} \leq v^*$.

Thus, with cost and transition structures suitably weighted to take into consideration the relationship between the quality of data provided by q and $\tilde{q}$, Proposition 9 is analogous to the value of information results found in ([45], [62]) and elsewhere.

We remark that Proposition 9 can provide lower and upper bounds on the value functions of any M-POMDP, in analogy to similar results found in [45] for the standard POMDP. For lower bounds on $v^\star$ and v_n, set $\xi(z'|\tilde{z}) = q(z'|\mu', \mu)$ for $\tilde{z} = \mu$, if $\tilde{q}(\tilde{z}|\mu', \mu) = 1$ if and only if $\tilde{z} = \mu$. Thus, if the first M-POMDP completely observes the modulation process, then its value functions are lower bounds on the value functions of any M-POMDP.

For upper bounds on $\tilde{v}$ and $\tilde{v}_n$, let $q(z'|\mu', \mu) = q(z')$ and set $\xi(z'|\tilde{z}) = q(z') = \xi(z')$. Then, the second M-POMDP gains no information about the state of the modulation process from the observation process (i.e., the modulation process is completely unobserved) and hence its value functions are upper bounds on the value functions of any M-POMDP.

The lower bound described above is by modifying the M-POMDP so that each observation received is equal to the modulation state. We can, additionally, demonstrate that direct knowledge of the underlying modulation process yields improved performance. Suppose we want to minimize the expected total discounted cost, where at each decision epoch the DM has available the information as in the M-POMDP, $\mathscr{I}_t$, but also knowledge of the modulation states $\{\mu_t, \ldots, \mu_1\}$. Feasible policies map $\mathscr{I}_t \cup \{\mu_t, \ldots, \mu_1\}$ into feasible actions at all epochs t. The DM is faced with a MDP defined by the operator $H_M : V_M \mapsto V_M$, where V_M is the space of bounded real-valued functions on $S \times M$,

$$H_M v(s, \mu) = \min_{a \in A(s)} \sum_{z', \mu'} P[z', \mu'|\mu] \left[c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a) v(s', \mu') \right].$$

We may view the lower bound generated in Proposition 10 as being generated by an *information relaxation*, à la [7]. Proof of the proposition follows by straightforward observation that all M-POMDP policies in Π are feasible for this MDP, but not all policies for this MDP are feasible for the M-POMDP. We remark that this bound may be improved by applying a

proper penalty term, akin to a Lagrangian relaxation, an idea developed in [7] and [38].

Proposition 10. $\sum_\mu x(\mu) v_M(s,\mu) \le v^*(s,x)$ *for all* $(s,x) \in S \times X$, *where* $v_M = H_M v_M$ *and* $v^* = Hv^*$.

2.4.4 Functional Description of Dynamics

In many applications, it is natural to describe the state dynamics on S via a function, $f : Z \times S \times A \mapsto S$ rather than transition probabilities. For example, in inventory problems with backlogging, $f(z', s, a) = s + a - z'$, where z' is the demand over t to $t+1$. Note that this is a special case for which the stochasticity of the transition probabilities is attributed solely to the observation process. For these applications, our operator H is equivalently defined

$$Hv(s,x) = \min_{a \in A(s)} \sum_{z'} \sigma(z'|x)\big[c(s,z',a) + \beta v\big(f(z',s,a), \lambda(z',x)\big)\big].$$

We show in the next proposition that if the transition probabilities satisfy a stochastic joint C3 property $\mathscr{P}^*_{\tilde{V}}$ with respect to $\tilde{V}$, then under the equivalent functional dynamic equation v composed with f exhibits the analogous joint C3 property, $\mathscr{P}^*$ (*e.g.* if $\mathscr{P}^*_{\tilde{V}}$ is jointly increasing on $S \times A$ with respect to $\tilde{V}$, then v composed with f is jointly increasing on $S \times A$).

Proposition 11. *Suppose that* $f : Z \times S \times A \mapsto S$ *is an equivalent functional description of state dynamics and* $v(\cdot, x) \in \tilde{V}$ *for all* $x \in X$. *Then,* $p(\cdot|z', \cdot, \cdot)$ *satisfies a stochastic joint C3 property* $\mathscr{P}^*_{\tilde{V}}$ *on* $S \times A$ *for all* $z' \in Z$ *so that for all* $k \in K$

$$\sum_{j \in J_k} \gamma_j p(\cdot|z', s_j, a_j) \precsim_{\tilde{V}} \sum_{i \in I_k} \gamma_i p(\cdot|z', s_i, a_i)$$

if and only if $v\big(f(z', \cdot, \cdot), \lambda(z', x)\big)$ *satisfies the joint C3 property* $\mathscr{P}^*$ *analogous to* $\mathscr{P}^*_{\tilde{V}}$ *for all* $(z', x) \in Z \times X$.

Now, we present sufficient conditions under which a structure on f is inherited by the

composition of v and f.

Proposition 12. *Suppose $f(z', \cdot, \cdot)$ satisfies a single point property $\mathscr{P}^*$ for all $z' \in Z$ with respect to some order on S, generated by the binary relation $\leq_S$, defined by convex weights $\{\gamma_j\}$, and $\tilde{V}$ is the space of bounded, non-decreasing, and convex functions on S. Then, for all $v \in V$ such that $v(\cdot, x) \in \tilde{V}$ for all $x \in X$, $v\big(f(z', \cdot, \cdot), \lambda(z', x)\big)$ has single point property $\mathscr{P}^*$ for all $(z', x) \in Z \times X$.*

2.4.5 Relative Optima & General State and Action Spaces

In all of the results included thus far, we have investigated structured policies which are optimal across the entire state space, S. For many applications, such as in inventory problems (as we will investigate below), this assumption may be overly restrictive. In fact, it may be the case that structured policies are only optimal with respect to some sequence of sets of states $\{S_t, t \geq 0\}$, for which structured actions are feasible. In such applications, we need an additional attainability condition which guarantees that the structured policy is always feasible (for an example in inventory applications, see Veinott's attainability condition in [56] and [57], and extended in [31]). The results above hold when the proposition conditions are appended with such an attainability condition.

Finally, while we present results here for the case where S and A are discrete, the results can be extended to hold for more general spaces, with measure theoretic considerations.

2.5 Extending MDP Structural Results

We now apply the results in Section 2.4 to a couple of important classes of MDPs. For each subsection we demonstrate that sufficient conditions guaranteeing the existence of an optimal structured policy for the MDP are sufficient for guaranteeing the existence of an optimal structured policy on S for the M-POMDP, modulated by the belief process $\{x_t, t \geq 0\}$. In Section 2.5.1 we discuss monotone optimal policies by generalizing conditions in [37]. We examine myopic policies in Section 2.5.2 and show that separability is a joint C3

property, adding separability to the list of C3 properties given by [46]. Thus, conditions from [47] which guarantee the existence of an optimal myopic policy for the MDP also guarantee the existence of an optimal myopic policy for the M-POMDP.

2.5.1 Monotone Optimal Policies

We now turn our attention to a motivating and important application of our results in Section 2.4 that will illustrate the manner in which our results may extend structural results in the MDP literature to the more robust modeling framework of the M-POMDP. Due to their appealing mathematical properties, various applications, and ease of implementation, there has been much interest in optimal policies for MDPs that are monotone ([63], [37], [43]).

We now present sufficient conditions for the existence of an optimal policy for the M-POMDP that is monotone on S, parametrized by belief state x. The proof of this proposition proceeds showing that B(a) - B(c) hold and applying Proposition 6. In the result below, we will use the terminology that transition probability functions satisfy some property "in the sense of first-order stochastic dominance", by which we mean that they satisfy the inequality test of satisfaction for that property with respect to $\precsim_{\tilde{V}}$ (as in Proposition 11), where $\tilde{V}$ is the space of real-valued non-decreasing functions on S.

Proposition 13. *Assume*

(i) $c(s, z', a)$ *non-increasing in* s *on* S *for all* $(z', a) \in Z \times A$

(ii) $p(\cdot|z', s, a)$ *stochastically non-increasing in* s *on* S*, in the sense of first-order stochastic dominance, for all* $(z', a) \in Z \times A$

(iii) $c(s, z', a)$ *subadditive in* (s, a) *on* $S \times A$*, for all* $z' \in Z$

(iv) $p(\cdot|z', s, a)$ *stochastically subadditive in* (s, a) *on* $S \times A$*, in the sense of first-order stochastic dominance, for all* $z' \in Z$.

Then, there exists an optimal value-policy function pair (v^*, π^*) *such that* $v^*(s, x)$ *is non-increasing in* s *on* S *for all* $x \in X$ *and* $\pi^*(s, x)$ *is non-decreasing in* s *on* S*, for all* $x \in X$.

2.5.2 Myopic Optimal Policies

We now consider myopic optimal policy structure, based on results in [47]. As in prior sections, we use our framework to extend the sufficient conditions given in [47] to the M-POMDP. We begin with a definition of separable functions.

Definition 6. *(Separability) A function* $f : S \times A \mapsto \mathbb{R}$ *is separable if there exists a function* $K : A \mapsto \mathbb{R}$ *and a function* $L : S \mapsto \mathbb{R}$ *such that* $f(s, a) = L(s) + K(a)$.

Proposition 14. *Suppose a function* $f : S \times A \mapsto \mathbb{R}$ *is separable, such that* $f(s, a) = K(a) + L(s)$. *Then, separability is a joint C3 property.*

Now, we present conditions under which the M-POMDP yields myopic optimal policies. Let $\mathscr{L}$ be the set of bounded, real-valued functions that map $S \times Z$ to $\mathbb{R}$, and let $\mathscr{K}$ be the set of bounded, real-valued functions that map A to $\mathbb{R}$.

Proposition 15. *Assume*

(i) $\exists K(z', \cdot) \in \mathscr{K}, L(\cdot, z') \in \mathscr{L} : c(s, z', a) = K(z', a) + L(s, z')$ *for all* $z' \in Z$

(ii) $p(\cdot|z', s, a)$ *is independent of* s

(iii) $a^*(x) \in \arg\min_{a \in A}\{G(x, a)\}$, *where*

$$G(x, a) = \sum_{z'} \sigma(z'|x) \left[K(x_t, a_t) + \beta \sum_{z''} \sigma\big(z''|\lambda(z', x)\big) \sum_{s'} p(s'|z', a) L(s', z'') \right]$$

(iv) $a^*(x_t)$ *is feasible for all* t

Then, the stationary deterministic policy $\pi^*(s, x) = a^*(x)$ *for all* $s \in S, x \in X$ *is optimal.*

2.6 Application

In this section, we consider the inventory control problem of [31]. The description of the problem setting and formulation is abbreviated, and we refer the reader to the original work for a more thorough and in-depth discussion. We include this brief description with the intention to demonstrate that this is an M-POMDP, and how the structural results found in this paper is consistent with the structural results from Sections 2.3 and 2.4. Namely, in Section 2.6.1, we re-derive the results from [31] which present conditions for a partially observed inventory control problem to have an optimal base-stock policy. The numerous applications in the literature review may be re-cast as M-POMDPs and structural results re-derived in a similar manner, justifying our claim that M-POMDPs model a broad class of important problems.

2.6.1 Inventory Control

We now model the inventory control problem considered in [31] as a M-POMDP. In this inventory problem, replenishment is assumed to be instantaneous and replenishment capacity is assumed to be infinite. Let $z_t = \big(z_{1,t}, z_{2,t}\big)$, where $\{z_{1,t}, t \geq 1\}$ is the demand process and $\{z_{2,t}, t \geq 1\}$ is the additional observation data (AOD) process. If the modulation process represents the underlying state of the economy, then the AOD process might provide macroeconomic data (e.g. consumer spending, housing starts), useful, together with demand data, for more accurately forecasting the state of the economy and, hence, future demand.

The single period cost accrued between epochs t and $t+1$ is $\hat{c}(z_{1,t+1}, y_t)$, where $y_t = s_t + a_t$. The function $\hat{c}(z_{1,t+1}, y_t)$ is assumed to be convex in y_t with $\lim_{|y_t|\to\infty} \hat{c}(z_{1,t+1}, y_t) = +\infty$. The single period cost function extensively considered in [31] is $\hat{c}(z', y) = p(z'_1 - y)^+ + h(y - z'_1)^+$, where $(\cdot)^+$ indicates the non-negative part, p is a shortage penalty per period for each unit of stockout, and h is a holding cost per period for each unit of excess inventory after

demand realization.

The dynamics of the inventory process are given by the functional equation $s_{t+1} = f(z_{1,t+1}, s_t, a_t)$, an example of which that is considered extensively in [31] is $f(z_{1,t+1}, s_t, a_t) = f(z_{1,t+1}, y_t) = y_t - z_{1,t+1}$. Thus, $p(s_{t+1}|s_t, z_{1,t+1}, a_t) = P[s_{t+1}|z_{1,t+1}, y_t] = P[s_{t+1} = f(z_{1,t+1}, y_t)]$.

Proposition 16. *Let $y^*(x) = \arg\min_y \sum_{z'} \sigma(z'|x)\hat{c}(z', y)$ for all $x \in X$, and suppose $y^*(x_t)$ is a feasible order-up-to level for all t. Then, the optimal value function, $v^*(s, x)$, is non-decreasing and convex in s for all $x \in X$, and an optimal policy is $\pi^*(s, x) = y^*(x) - s$.*

We note that the condition $y^*(x_t)$ is a feasible order-up-to level for all t is an attainability condition. In [31], this feasibility condition would be replaced by an extension of Veinott's attainability condition in [56] and [57].

2.7 Computational Procedures

In this section, we detail ways in which the M-POMDP conditioning structure (equations (2.2) and (2.3)) admits to approximate computational solution procedures that are special to M-POMDPs.

POMDPs, fundamentally, are difficult to solve for large instances due to the fact that the belief space X contains an uncountably infinite number of possible belief vectors. There have been various approaches in the literature that seek to overcome this issue, which may be broadly characterized as exact methods, fixed-grid approximation methods, and belief trajectory simulation methods.

Exact methods are based upon the value iteration algorithm and seek to solve the POMDP exactly by utilizing the piecewise linear and concave structure of the value function with respect to x to construct the defining facets of v. [48] and [45] were the first to take this approach in their seminal papers. [25] improved upon the complexity of this approach by utilizing linear programming to construct the facet vectors.

The most well-known fixed grid method is due to [30], in which a Freudenthal triangulation of the belief space generates point-based value function estimates at grid points uniformly dispersed across X. Another similar approach is to generate a random grid on X, which is shown to perform similarly to the Freudenthal triangulation method, as in [22]. Neither of these approaches utilize any of the information inherent to the dynamics of the belief process.

Belief trajectory simulation methods are based upon the intuition that, for many problems, there are only a small subset of beliefs that are *reachable* under an optimal policy. Various approaches in the literature successively build a grid on X by alternating at each epoch between sampling new beliefs and performing value iteration operations on the new belief states ([34], [49]).

We first present an *a priori* belief trajectory simulation method for constructing a discrete grid approximation, $X' \subset X$, which utilizes the *actual dynamics* of the modulation and observation processes, while alleviating the computational burden associated with past approaches due to the fact that learning in M-POMDPs is *passive* and independent of control. This method turns solving the M-POMDP into solving a completely-observed MDP with state space $S \times X'$. Then we present a *real-time* heuristic method that utilizes the M-POMDP structure to find approximate solutions for each belief state, as they are encountered in real-time. Finally, we discuss an information relaxation approach based on the lower bound in Proposition 10.

2.7.1 Simulation-based Approximation Method

Suppose we have a metric space $(X, \|\cdot\|)$, where $\|\cdot\|$ is the sup-norm and X is the belief space. Let $X_d \triangleq \left\{x \in X : \exists x' \in X : x = \frac{\lfloor x' \cdot 10^d \rfloor}{10^d}\right\}$, the grid of points in X rounded to the d-th digit. Note that $X_d \subset X$. We detail the so-called X' solution procedure for M-POMDPs.

In step 0, we initialize the solution procedure. We note that d should be a positive integer and controls the fineness of the grid. The cardinality parameter, K, determines how

0. *Initialization.* Initialize belief distribution, modulation state, number of simulation runs, mesh parameter, and cardinality parameter — x_0, μ_0, N, d, and K respectively.

1. *Belief simulation.* Generate, according to $P[z', \mu|\mu]$ the sequences $\{z_t, t = 1, \ldots, N\}$ and $\{\mu_t, t = 0, \ldots, N\}$. Then compute recursively $\{x_t, t = 1, \ldots, N\}$ such that $x_{t+1} = \lambda(z_{t+1}, x_t)$ for $t = 0, \ldots, N-1$.

2. X' *definition.* Let $\tilde{x}_t$ be x_t rounded to the d-th digit and let $X' \triangleq \bigcup_{i=1}^{K} \tilde{x}_{(i)}$, the K-th most frequently visited balls of radius 10^{-d} in X.

3. *Solving the MDP with state space* $S \times X'$. Solve the modified completely observed MDP with optimality equation

$$\hat{v}(s,x) = \min_{a \in \mathcal{A}(s)} \sum_{z'} \sigma(z'|x) \Bigg[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) \hat{v}\Big(s', x'(z',x)\Big) \Bigg],$$

where $x'(z',x) \approx \lambda(z',x)$ and $x'(z',x) \in X'$.

Figure 2.2: The X' method.

many points will be included in the approximate grid.

In step 1, we simulate a trajectory of the beliefs by simulating the evolution of observations and modulation states according to the underlying Markov chain governing the dynamics, and recursively performing the belief update operations according to these observations and modulation states. So long as the Markov chain for the modulation states is ergodic, simulating one long trajectory should be sufficient for approximating a steady state distribution of modulation states. We note that this step is simulating a *passive* learning environment since the belief updates are independent of control under the M-POMDP conditioning assumptions, guaranteeing that the learning operation for M-POMDPs is computationally tractable.

In step 2, we determine $\{\tilde{x}_t, t = 0, \ldots, N\}$, the set of simulated belief states rounded to the d-th digit, so that $\tilde{x}_t$ is the unique point in X_d such that x_t is within a ball of radius 10^{-d} of $\tilde{x}_t$. Let $\tilde{X} = \bigcup_{t=1}^{N} \tilde{x}_t$. (Note that $\tilde{X} \subset X_d$.) There is a complete order on $\tilde{X}$ induced by the

binary operator, $\preceq$, defined so that

$$\tilde{x}_{(i)} \preceq \tilde{x}_{(j)} \Leftrightarrow i < j \text{ and } \sum_{t=1}^{N} \mathbf{1}\left\{\left\|x_t - \tilde{x}_{(i)}\right\| \leq 10^{-d}\right\} \leq \sum_{t=1}^{N} \mathbf{1}\left\{\left\|x_t - \tilde{x}_{(j)}\right\| \leq 10^{-d}\right\}.$$

This order counts the number of simulated beliefs that are rounded to a particular $\tilde{x}$ and ranks them. We then define X' to be the K-th most frequently visited rounded beliefs (Note that $X' \subset \tilde{X} \subset X_d \subset X$). Of course, X' has cardinality K, so it is finite in dimension.

Finally, in step 3 we are left with the M-POMDP optimality equation, below

$$v(s,x) = \min_{a \in \mathcal{A}(s)} \sum_{z'} \sigma(z'|x) \left[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) v\big(s', \lambda(z',x)\big) \right], \quad \forall (s,x) \in S \times X'.$$

Our remaining challenge is that $\lambda(z',x)$ may not be in X' for a given (z',x). Suppose $x \in X'$. The hope is that $\exists x'(z',x) \in X'$ such that $\lambda(z',x) \approx x'(z',x)$, and that $v(\cdot, \lambda(z',x)) \approx v(\cdot, x'(z',x))$. These assumptions may not hold if either $\lambda(z',x)$ is not near any point in X' (although intuitively, in most cases, it should be since we chose X' on the basis of frequently visited belief vectors in our simulation), or if $\lambda(z',x)$ is near a facet of the Sondik regions of X, so that $v(\cdot, x'(z',x))$ is not a good approximation to $v(\cdot, \lambda(z',x))$. There are many ways we could define $x'(z',x)$, such as $x'(z',x) \triangleq \arg\min_{x' \in X'} \{\|x' - \lambda(z',x)\|\}$.

This creates a well-defined MDP, with state space $S \times X'$, which serves as our approximate model for the M-POMDP. The benefits of this method is that we reduce drastically the number of possible belief states that we need to consider in the M-POMDP by using the *actual dynamics* of the system, which makes it better-suited than uniform or random grid methods for each particular problem instance ([30], [22]).

2.7.2 Heuristic Solution Procedure

We now present an alternative, heuristic solution procedure that must be implemented in an online manner. The fundamental idea is to map the M-POMDP into a related completely

observed MDP with a state space on $S \times Z$ rather than on $S \times X$. We may assume that Z is finite in its cardinality, and thus this mapping is a state space dimensionality reduction technique (as is the X' procedure, above). The tradeoff is that we must solve such an MDP at each time epoch in order to capture the belief dynamics.

---

0. *Initialization.* Assume (s_0, x_0) is given. Set $t = 0$.
1. Solve the completely observed MDP for all (s, z'):

$$v'_{z'}(s, x_t) = \min_{a \in A(s)} \left\{ c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a) \sum_{z''} \sigma(z''|\lambda(z', x_t)) v'_{z''}(s', x_t) \right\}.$$

   Let $\delta^*_{z'}(s, x_t)$ be an optimal policy, mapping $S \times Z$ into A.
2. Choose action a_t to equal $\delta^*_{z'}(s_t, x_t)$ with probability $\sigma(z'|x_t)$.
3. Observe the observation z_{t+1} (which will equal z' with probability $\sigma(z'|x_t)$). Set $x_{t+1} = \lambda(z_{t+1}, x_t)$.
4. Observe the state s_{t+1} (which will equal s' with probability $p(s'|z_{t+1}, s_t, a_t)$).
5. Increment $t \leftarrow t + 1$; go to 1.

Figure 2.3: Real-time heuristic method.

The intuition behind the procedure begins with the observation of the following inequality

$$\min_{a \in A(s)} \sum_{z'} \sigma(z'|x) h\big(s, a, v(\cdot, \lambda(z', x))\big) \geq \sum_{z'} \sigma(z'|x) \min_{a \in A(s)} h\big(s, a, v(\cdot, \lambda(z', x))\big).$$

By pulling the minimization inside the summation, the idea is to establish a lower bound on v^* by solving a related problem. We formalize this intuition in the subsequent proposition. Let

$$\tilde{H}_{z'} \tilde{v}(s, x) = \min_{a \in A(s)} \left\{ c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a) \sum_{z''} \sigma(z''|\lambda(z', x)) \tilde{v}_{z''}\big(s', \lambda(z', x)\big) \right\},$$

and let $\tilde{v}_{z'}$ be the unique fixed point of $\tilde{H}_{z'}$.

Proposition 17. $v^*(s,x) \geq \sum_{z'} \sigma(z'|x)\tilde{v}_{z'}(s,x)$, *for all* $(s,x) \in S \times X$.

Solving for $\{\tilde{v}_{z'} : z \in Z\}$ is no more computationally tractable than solving for v^* due to the cardinality of X and the dependence of $\tilde{v}_{z'}$ on $\lambda(z',x)$. In developing our heuristic procedure, we seek an approximation to $\{\tilde{v}_{z'} : z \in Z\}$ for a fixed x. If we assume $\max_{z'} \|x - \lambda(z',x)\|$ is small, then it is reasonable to assume that $\tilde{v}_{z''}(s', \lambda(z',x))$ is close to $\tilde{v}_{z''}(s',x)$ in many cases. This is effectively a *learning rate* assumption (that learning is incremental and gradual), and is one that has been made in the literature, *e.g.* [31]. We then define a completely observed MDP with state space $S \times Z$:

$$v'_{z'}(s,x) = \min_{a \in A(s)} \left\{ c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) \sum_{z''} \sigma(z''|\lambda(z',x)) v'_{z''}(s',x) \right\} \tag{2.5}$$

This is the intuition behind step 2 in Figure 2.3. Since this approximation is for a fixed x, it is amenable to an online implementation, where this completely observed MDP is solved for each x_t.

We remark that the following is *likely* to be a valid inequality (although not necessarily)

$$v^*(s,x) \geq \sum_{z'} \sigma(z'|x) v'_{z'}(s,x),$$

where $v'_{z'}$ is the fixed point of Equation 2.5. We use Equation 2.5 to develop a heuristic that, for a given (s,x), chooses action $\delta^*_{z'}(s,x)$ (an optimal policy mapping $S \times Z$ into A, for this approximate MDP) with probability $\sigma(z'|x)$. This randomized policy is a *probability matching* heuristic.

2.7.3 Information Relaxation

In Proposition 10, we establish that there always exists an information relaxation lower bound by solving the MDP generated by H_M. Of course, if we have available a heuristic

policy, π_h, then the value of that heuristic v^{π_h}, which may be found by solving for the fixed point of H^{π_h} either exactly or by Monte Carlo simulation, is an upper bound on v^*. Thus, if the difference $v^{\pi_h} - v_M$ is small, then the heuristic π_h is approximately optimal.

This method of utilizing the information relaxation lower bound might be combined with some of the other solution procedures in this section to generate π_h. For example, if π_M (a function from $S \times M$ to A) is optimal for the information relaxation MDP generated by H_M, then we might consider π_h to be the randomized policy such that $\pi_h(s, x) = \pi_M(s, \mu)$ with probability $x(\mu)$. If $\pi_{X'}$ is the optimal policy generated for the MDP in Step 3 of Figure 2 (a function from $S \times X'$ to A), then one might consider $\pi_h(s, x) = \pi_{X'}(s, \bar{x})$, where $\bar{x} = \arg\min_{x' \in X'} \|x - x'\|$.

2.7.4 Tradeoffs & Remarks

There are many ways in which we may craft approximate solution procedures that utilize the M-POMDP model structure. The procedures presented here are not meant to be exhaustive, but rather illustrative examples of how the M-POMDP model structure might admit to solution procedures that are unique to the M-POMDP.

In comparing the two solution procedures in Sections 2.7.1 and 2.7.2, there are tradeoffs that may make one or the other more well-suited for particular problem instances. For the simulation-based approach, it may be the case that even though the number of reachable belief vectors is finite, it may be very large. Thus the cardinality parameter K would need to be very large in order to guarantee that the MDP on $S \times X'$ is a good approximation to the M-POMDP. In such cases, if the cardinality of Z is small and much less than the number of reachable belief vectors, then the real-time heuristic procedure may be much more tractable to implement, even though the MDP on $S \times Z$ must be solved at each decision epoch.

We recall that the number of operations per successive approximations step of an MDP is on the order of the cardinality of the state space squared times the cardinality of the action space. The cardinality of the state space for the simulation-based approximation method is

K times the number of permissible actions, whereas this cardinality is $|Z|$ times the number of permissible actions for the heuristic solution procedure, where $|Z|$ is the cardinality of the set Z. Thus, the simulation-based method requires no more than the number of operations per successive approximations step than the heuristic solution procedure if and only if $K < |Z|$. We recall, however, that that the heuristic solution procedure requires its MDP to be solved at each decision epoch.

Finally, we note that the procedures we detail in this section are not straightforwardly applicable to the general POMDP. The X' method in Section 2.7.1 utilizes the passive learning property of M-POMDPs, which is not a characteristic of general POMDPs. If we extend the real-time heuristic procedure in Section 2.7.2 to the general POMDP, it generates the best possible myopic policy, which can be decidely sub-optimal in many situations. Lastly, the information relaxation approach in Section 2.7.3 utilizes the MDP generated by H_M, which is not applicable as a dual problem to the general POMDP.

2.8 Conclusions

We have presented and analyzed the M-POMDP, a specially structured POMDP that explicitly considers factors that affect the decision-making environment, are not controllable by the DM, and are partially observed by a data-driven observation process. We demonstrated that there is a broad class of decision-making environments with these salient characteristics, and examples of this class have been considered previously in the literature, but without a unifying analytical framework. We demonstrated that the characteristics of these decision-making environments lead to interesting properties for the M-POMDP; *e.g.*

1. They model a broad class of problems.

2. They inherit the structural characteristics of the value function and optimal policy function from their MDP analogs.

3. They inherit concavity of the value function from general POMDPs. With cost and

transition structures suitably weighted to take into consideration the relationship between the quality of data, better observation quality improves optimal system performance.

4. They separate at each decision epoch the task of learning the underlying modulation state and optimizing, leading to improved tractability of solution procedures that are special to the M-POMDP.

CHAPTER 3
THE VALUE OF INFORMATION AND SUPPLY CHAIN AGILITY IN MANAGING DEMAND UNCERTAINTY IN INVENTORY SYSTEMS

3.1 Introduction

Demand uncertainty can have detrimental and propagating effects in a supply chain if improperly managed. These effects include cost pressures due to a supply-demand imbalance, either incurring undue inventory holding, production, and distribution costs due to overly optimistic demand expectations, or the opportunity costs of stock-outs and customer attrition due to supply shortages and poor operating policy.

The big data revolution in the modern economy provides firms with unprecedented availability of data, and thus potential for addressing demand uncertainty through data-driven market insights and demand forecasts. Firms feel pressure to invest in an information infrastructure — data access and engineering, IT systems, and quantitative talent — in order to compete. At the same time, as the speed of information transmittal increases, consumer trends are putting pressure on supply chains to cut cost and be more agile — to deliver products or services more quickly. Creating agility, whether through a redesign of the supply chain network or product architecture, allows the firm to react and quickly respond to changing demand, reducing the burden of projecting demand too far into the future.

In this paper, we consider the effects of capital investments in supply chain agility and an improved information infrastructure on an inventory control problem that explicitly incorporates three sources of demand uncertainty — competitive pressures and the state of the macroeconomy, product-specific demand variability, and information quality. The decision-maker (DM) must make inventory ordering decisions on the basis of multiple information sources — historically observed demand and noise-corrupted observations of an

exogenous modulation process that models broad macroeconomic and competitive effects. Further ordering decisions take time to manifest in the market, *e.g.* due to transportation, distribution, and/or production time, so the DM must consider how ordering decisions will realize in potential evolutions of the market, demand, and observational data.

We prove conditions under which, for a fixed lead time and information quality, an optimal policy for the DM has a base stock structure, in which the base stock levels are with respect to the inventory position, *i.e.* total pipeline inventory level, and parametrized by the DM's Bayesian belief distribution over the possible states of the macroeconomy. Even with this base stock structure, this optimal policy can be difficult to compute due to cardinality of the Bayesian belief simplex and the number of potential evolutions of the market, demand, and observations over the lead time between when ordering decisions are made and their realization. In this paper, we present a computationally efficient method for generating these base stock levels. We show that the optimal base stock levels are constant within regions of the belief simplex defined by two linear inequalities. We then show how to approximate the optimal base stock levels by generating a linear partition of the Bayesian belief simplex using a fixed grid approximation to the simplex, simulating trajectories of future observational uncertainties, and constructing the partitioning hyperplanes using successive support vector machines.

Finally, we investigate how changes in information quality and supply chain agility impact optimal inventory policies and system performance. This investigation is motivated by the following questions:

- What is the *value of information* and the *value of agility* in managing demand uncertainty? What are the limitations?
- How can a manager *quantify*, or *measure*, the effects of their operating policies and strategic decisions on business objectives?
- Should investment money be allocated towards better data access/quality and infor-

mation processing, or to a more agile product/supply chain design?

In considering these questions, we prove that better information — *e.g.* in the form of better forecasts, an improved IT infrastructure, expanded data access — defined in terms of signals passed through Markov noisy channels, improves long run costs under an optimal policy (but not necessarily under a sub-optimal policy). Further, we provide conditions on the firm's cost structure under which these effects hold for investments in a more agile supply chain, thereby proving both the value of information and agility. Finally, we show how to evaluate numerically, using Monte Carlo simulation and regression, the effects of investments in better quality information and agility in terms of discounted costs, stock-outs, and risk of high-cost scenarios. We further discuss how these evaluations may serve as the basis of ROI calculations, or an optimization model for determining capital allocation.

3.1.1 Literature Review

This research brings together ideas from various strands of literature.

Inventory Control. In this paper, we consider a model that generalizes approaches in the Markov-modulated demand and Bayesian updating inventory control literatures. [31] consider a completely observed inventory control problem with instantaneous replenishment and backlogging, augmented with a partially observed exogenous modulation process that models the underlying state of the economy. This modulation process is observed by demand realization and an "all other data" (AOD) process. This is, to the best of our knowledge, the only inventory control model in the literature to explicitly incorporate a generalized source of macroeconomic data, and is the modeling framework that most resembles our own. They prove the optimality of base stock policies that are parametrized by the Bayesian belief distribution over the modulation space, and that these base stock levels are constant within regions of the belief space defined by a finite set of linear inequalities. We recover similar results in this paper, but whereas [31] assume instantaneous replenishment, we consider ordering decisions with delayed replenishment, a generalization that we

also consider to be more realistic in practice. Further, we present a way to incorporate support vector machines, a popular technique for classification and regression problems in machine learning, in order to efficiently construct the belief space partition.

[54] consider a completely observed inventory control model with a Markov-modulated demand environment, in which the modulation process is partially observed by demand realizations. They determine the optimality of base stock policies that are parametrized by a Bayesian belief distribution over the modulation space. However, due to the intractability of solving for base stock levels at each belief vector they present and evaluate multiple heuristics in order to determine the quality of each heuristic in various demand environments. We note that we consider additional sources of information explicitly in this paper and utilize the problem structure to avoid computational intractabilities due to the cardinality of the belief simplex. [7] apply information relaxation duality to determine tight bounds on the performance of the heuristics considered in [54].

POMDPs. [5] analyze specially structured POMDPs, so-called modulated POMDPs or M-POMDPs, in which there are multiple levels of effects — a completely observed and controllable state process weakly coupled through observations to a partially observed and uncontrollable modulation process. For this class of models, value function and optimal policy structure are inherited from analogous MDPs and feature a passive learning environment. Our model satisfies assumptions on the dynamics as in the M-POMDP, and may be properly reformulated into this framework. We prove the optimality of base stock policies, which is a structure inherited from a simpler completely observed inventory control model. Additionally, our model considers time-lagged actions, the effects of which to our knowledge have not been analyzed for general POMDPs. [3] investigate POMDPs with time-lagged and noise corrupted observations, but not actions.

Value of Information. The focus of this paper is closely related to the *value of information*, which is a loosely and variously defined concept in the literature. [62] show for POMDPs that worse observation quality, defined in terms of information corrupted by a

Markov noisy channel, degrades system performance under an optimal control policy, but may not degrade performance under suboptimal policies. We adopt this notion of observation quality in our investigation of the value of AOD information. [19] and [27] consider the value of information flow between a supplier and retailer under different contexts, defined in terms of expected costs, that resemble in principle our focus in its application to supply chains, but are significantly different in approach and results. We note that often the value of information is defined in terms of expected costs, but we additionally analyze the effects of managerial decisions on other operational metrics, such as variance in Section 3.5 and stock-outs in Appendix B.5.

Support Vector Machines. As we have mentioned previously, we show that the structural properties of our problem admit to an optimal policy dependent upon a partition of the space of Bayesian belief distributions over the modulation states. We introduce a sequential procedure for constructing the partition utilizing support vector machines, a popular machine learning technique introduced in [14] for classification problems.

Hidden Markov Models. The modulation, AOD, and demand processes in our model may be viewed as a hidden Markov model (HMM). Thus, our inventory control model may be considered a MDP weakly coupled to a HMM via the demand process. Methods for determining the underlying HMM primitives in our model are outside the scope of this paper, but for a survey of such techniques and HMM theory we refer the reader to [16].

Hierarchical/Multilevel Modeling & Separation. Our modeling approach in this paper is analogous to developments in the statistical literature pertaining to hierarchical, or multilevel, modeling ([20]). In statistics, this refers to nested or composed statistical models, in which compositions represent dependencies. Every super-model is dependent upon effects modeled in the sub-model, but the sub-model is *not* dependent upon effects in the super-model. This is also similar to the separation principle in the control literature, in which state estimation is separated and performed first, and then optimization follows given the results of state estimation ([6], [55]). Thus, state estimation is the sub-model and optimiza-

tion is the super-model. In our modeling framework (from the lowest level model to the highest), the dynamics of the macroeconomy are modeled as a Markov chain and called the *modulation process*. The demand and AOD information processes are modeled as a Hidden Markov Model (HMM) in which the latent stochastic model is the modulation process. The inventory control model is a M-POMDP, which takes as a constituent sub-model the demand/information/modulation HMM. In other words, the M-POMDP is composed of the HMM, which in turn is composed of the Markov chain. There is a natural hierarchy to this problem, in which the highest level model corresponds to the most local effects to the firm (inventory ordering decisions) and the lowest level model corresponds to the least local effects to the firm (the macroeconomy). The M-POMDP regards the firm's inventory decisions and interaction with the economic environment; the HMM models how the macroeconomy generates information; the Markov chain models how the macroeconomy evolves.

3.1.2 Outline

The rest of the paper is structured as follows. In Section 3.2, we present the initial model formulation and discuss how the model primitives explicitly reflect three types of demand uncertainty. In Section 3.3, we show how the initial model formulation can be alternatively expressed into a form that is more conducive to structural analysis. Section 3.4 is dedicated to proving the optimality of base stock policies and how to efficiently compute the order-up-to levels using the structure of the base stock levels with respect to the Bayesian belief simplex and support vector machines. In Section 3.5, we consider how/whether capital should be allocated to manage demand uncertainty through better information or supply chain agility. In subsections 3.5.1 and 3.5.2, we prove how the value function is affected by AOD information quality and lead time. In subsections 3.5.3 and 3.5.4, we perform a Monte Carlo numerical analysis and regression sensitivity analysis on a numerical example in order to demonstrate how to quantify the effects of operational performance with respect

to information quality and agility. Then, we discuss how this numerical output might be embedded in a capital allocation decision-making process. We conclude and discuss future research directions in Section 3.6.

3.2 Problem Formulation

We consider a discrete-time stochastic dynamic program with the following constituent processes:

- $\{s_t, t = 0, 1, \ldots\}$ is defined to be the *inventory level process*, where s_t is the inventory level at the decision epoch t prior to satisfying demand and being replenished.
- $\{d_t, t = 0, 1, \ldots\}$ is defined to be the *demand process*, where d_t is the demand that becomes known just before decision epoch t.
- $\{a_t, t = 0, 1, \ldots\}$ is the *replenishment process*, where a_t is the replenishment decision made at decision epoch t.
- $\{z_t, t = 1, 2, \ldots\}$ is the *additional observation data (AOD) process*, where z_t represents data that becomes known just before epoch t from sources in addition to demand that might be useful in more accurately forecasting demand. The set of all possible observations is Z and is assumed to be finite.
- $\{\mu_t, t = 0, 1, \ldots\}$ is the *modulation process*, models the underlying and perhaps only partially observed forces that affect demand but are not controllable. The set of all modulation states is $M = \{\mu_1, \ldots, \mu_{|M|}\}$ and is assumed to be finite.

Further, we assume that the inventory, demand, and replenishment processes are related through the stochastic difference equation $s_{t+1} = s_t + a_{t-\tau} - d_t$, which assumes backlogging is allowed, where τ is the replenishment delay. This equation can be described as a conditional probability $P[s_{t+1}|s_t, d_t, a_{t-\tau}]$. We remark that this definition of inventory dynamics

differs from the usual definition in that we allow the decision maker to know demand at epoch t, whereas the typical formulation assumes that d_t represents the (random variable) demand realized *between* epochs t and $t+1$. We further remark that when $\tau = 0$, the replenishment decision is made knowing the current inventory level and the current number of orders to be fulfilled, thus resembling a build-to-order (BTO) production environment. Hence, this model considers both BTO when $\tau = 0$ and more traditional production environments when $\tau > 0$.

The modulation process is assumed to be related to the demand and AOD processes by the conditional probability $P[z_{t+1}, d_{t+1}|\mu_{t+1}, \mu_t]$ and has dynamics described by the transition probability $P[\mu_{t+1}|\mu_t]$. We assume the dynamics are described via the following conditional probabilities,

$$P[z_{t+1}, s_{t+1}, d_{t+1}, \mu_{t+1}|s_t, d_t, \mu_t, a_{t-\tau}] = P[s_{t+1}|s_t, d_t, a_{t-\tau}]\cdot P[z_{t+1}, d_{t+1}|\mu_{t+1}, \mu_t]\cdot P[\mu_{t+1}|\mu_t]. \tag{3.1}$$

While it is not necessary for most of the results we present in this paper, we will often assume for simplicity that the demand and AOD processes are conditionally independent, *i.e.* $P[z_{t+1}, d_{t+1}|\mu_{t+1}, \mu_t] = P[z_{t+1}|\mu_{t+1}, \mu_t] \cdot P[d_{t+1}|\mu_{t+1}, \mu_t]$. We will also assume that the support of the demand distribution $P[d_{t+1}|\mu_{t+1}, \mu_t]$ is finite for all μ_{t+1}, μ_t.

We denote the single-period cost accrued at epoch t as $C_\tau(s_t, d_t, a_t, a_{t-\tau}) = c_\tau a_t + h_\tau(s_t + a_{t-\tau} - d_t)^+ + p_\tau(d_t - s_t - a_{t-\tau})^+$, where $(x)^+ = \max(0, x)$. The per-unit holding cost is h_τ, the per-unit purchase cost is c_τ, and p_τ is the per-unit underage cost. For notational simplicity, we'll suppress the τ from p and c in the formulations below when it is clear to which τ we are referring.

At decision epoch t, the decision-maker (DM) makes decisions on the basis of their information pattern, $\mathscr{I}_t = \{s_t, \ldots, s_0, d_t, \ldots, d_0, z_t, \ldots, z_1, a_{t-1}, \ldots a_{-\tau}\}$. A feasible policy, π, maps $\mathscr{I}_t$ into a replenishment decision. Let Π be the space of feasible policies. The overall

objective is to minimize the total expected discounted costs across the infinite horizon,

$$\min_{\pi \epsilon \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t C_\tau(s_t, d_t, a_t, a_{t-\tau}) | \mathscr{I}_0\right],$$

where $0 < \beta < 1$ is the discount factor, and $a_t = \pi(\mathscr{I}_t)$ for all t.

The optimality equation is $v = Hv$, where

$$Hv(\mathscr{I}_t) = \min_{a_t} \left\{C_\tau(s_t, d_t, a_t, a_{t-\tau}) + \beta \mathbb{E}\left[v(\mathscr{I}_{t+1}) | \mathscr{I}_t, a_t\right]\right\}. \tag{3.2}$$

By results in [37], H is a contraction operator and there exists a unique fixed point v^* of H, such that $v^* = Hv^*$, representing the total expected discounted costs accrued by an optimal policy. This fixed point can be attained by value iteration — repeated application of H, so that $\lim_{n\to\infty} \|Hv_n - v_n\| = 0$, where $\|\cdot\|$ is the sup-norm and $v_{n+1} = Hv_n$.

We note, additionally, that there are many alternative problem formulations that one might consider, for which the general modeling framework and line of reasoning in this paper would hold. For instance, in this paper we generate policies with respect to discounted costs, and later incorporate other metrics in the evaluation scheme. We may be interested in explicitly considering risk measures into the problem formulation. Other alternative formulations might consider pricing or promotions along with inventory management, or assume lost-sales for stock-outs.

3.2.1 Types of Demand Uncertainty

In this subsection we relate the model primitives to their applied context. In our model, there are three explicit sources of demand uncertainty: the state of the economy, product-specific demand variability, and information quality. We investigate in this paper how to make optimal (or near-optimal) replenishment decisions in the presence of these sources of demand uncertainty, and two primary ways of mitigating these uncertainties — (1) reducing lead times by making the supply chain more agile, and (2) better data or information pro-

cessing. (In Appendix B.5, we additionally consider artificially adjusting the cost structure in order to generate policies that are robust to specific uncertainties.)

The state of the economy. The first source of demand uncertainty is the state of the economy. Every firm is impacted by forces that they cannot control, including competitive pressures, macroeconomic trends, financial markets, consumer sentiment, *etc.* In our model, these exogenous forces are modeled through the modulation process, $\{\mu_t\}$. In the Markov-modulated demand literature, the modulation process is variously described as the "core" process ([2]), the "world" process ([64]), or the "state of the economy" ([31]). In our context, each μ_t is supposed to capture a possible "state of the economy", and the stochastic process modeling the evolution of the state of the economy is assumed to be a stationary Markov chain. Thus, the dynamics and inherent uncertainties as to these uncontrollable forces are modeled through the conditional probabilities $P[\mu_{t+1}|\mu_t]$ and a stochastic matrix $\mathcal{M}$ such that $\mathcal{M}[i,j] = P[\mu_{t+1} = \mu_j|\mu_t = \mu_i]$. We assume that the definitions of each modulation state and the Markov transition probabilities may be constructed or estimated by macroeconomists considering the confluence of myriad market effects. The nature of these transition probabilities impact the DM's objective through the dependence of the demand process on the modulation process. Thus at time t, the uncertainty as to the next state of the economy, μ_{t+1}, yields additional uncertainty as to the realization of future demand. This matches our intuition that demand forecasting requires considering explicitly the changing market.

Product-specific demand variability. A second source of demand uncertainty is product-specific demand variability — that is, the uncertainty in demand due to the nature of the product itself, with all other exogenous factors held constant. In our context, this type of demand uncertainty is modeled by the conditional probabilities, $P[d_t|\mu_{t+1}, \mu_t]$, so we might view each modulation transition (μ_t, μ_{t+1}) as representing a particular *demand model*, $d_t|\mu_{t+1}, \mu_t$.

Information quality. The AOD process may be considered in various ways. One way is

considering each AOD observation, z_t, to represent data that are collected and informs the DM about the state of the economy, such as SEC filings, housing starts, *etc.* Additionally, we might consider z_t to be the output of some firm's internal information processing mechanism. For instance, it might be that z_t is a demand forecast, or an estimation of the state of the economy, that is the end result of an analytical process. Thus, the *quality* of these AOD observations may reflect the degree of access to good data, the level of quantitative talent inside the firm, the quality of IT systems, or any combination of these factors.

3.3 Preliminary Results

The optimality equation (3.2) is not suitable for solving due to the fact that $\mathscr{I}_t$ grows linearly in t, so we must reformulate the problem. For the case when $\tau = 0$, results in [45] and [48] guarantee that (s_t, d_t, x_t) is a sufficient statistic for optimal control, where $x_t = \{x_t(\mu) : \mu \in M\}$ is the *belief vector* such that $x_t(\mu) = P[\mu_t = \mu | \mathscr{I}_t]$. We denote by X the space of possible probability mass functions over M, the *belief simplex*, where

$$X \triangleq \left\{ x \in \mathbb{R}^{|M|} : x(\mu_j) \geq 0, j = 1, \ldots, |M|, \quad \sum_{j=1}^{|M|} x(\mu_j) = 1 \right\}.$$

[3] extend these results to POMDPs with delayed and noise-corrupted state observations. The following proposition is similar to these results and establishes a sufficient statistic for our problem setting in which ordering decisions are delayed τ decision epochs between placement of the order and fulfillment.

Proposition 18. *The vector* $(s_t, d_t, a_{t-1}, \ldots, a_{t-\tau}, x_t)$ *is a sufficient statistic for optimal control.*

Thus, we may reformulate the problem around the sufficient statistic $(s_t, d_t, a_{t-1}, \ldots, a_{t-\tau}, x_t)$. It will be convenient to use the notation $(s, d, a_{-1}, \ldots, a_{-\tau}, x)$ to indicate the sufficient statistic $(s_t, d_t, a_{t-1}, \ldots, a_{t-\tau}, x_t)$ (suppressing the t for notational simplicity) and (d', z', μ', s') to represent $(d_{t+1}, z_{t+1}, \mu_{t+1}, s_{t+1})$. The optimality equation then becomes

$v = H^{(a)}v$ (the (a) superscript is an index defining the first reformulation), where $H^{(a)}$ is defined as

$$H^{(a)}v(s,d,a_{-1},\ldots,a_{-\tau},x) = \min_{a}\Bigg\{C_{\tau}(s,d,a,a_{-\tau}) + \beta\sum_{d',z'}\sigma(d',z'|x)v\big(s-d+a_{-\tau},d',a,\ldots,a_{-\tau+1},\lambda(d',z',x)\big)\Bigg\}, \tag{3.3}$$

and where

$$\sigma(d',z'|x) = \sum_{\mu',\mu} P[d',z'|\mu',\mu]P[\mu'|\mu]x(\mu)$$
$$\lambda(\mu'|d',z',x) = \frac{\sum_{\mu} P[d',z'|\mu',\mu]P[\mu'|\mu]x(\mu)}{\sigma(d',z'|x)},$$

provided $\sigma(d',z'|x) > 0$. The vector $\lambda(d',z'|x)$ is the Bayesian posterior belief vector over the modulation space M, given that the prior belief vector is $x_t = x$, the DM observed demand $d_{t+1} = d'$, and the AOD observation is $z_{t+1} = z'$. Similarly, $\sigma(d',z'|x)$ represents the probability of observing demand d' and AOD observation z' given the prior belief vector x. We note that the posterior belief update, λ, is independent of control and thus we have a *passive learning* environment. From a technical perspective, the passive nature of the learning environment is due to the independence of the modulation state dynamics of control, as is discussed in [5]. In our applied setting, this reflects our assumption that the modulation process models forces that the DM must consider but cannot control.

We may further refine the optimality equation by defining $y_t = s_t + \sum_{j=1}^{\tau} a_{t-j} + a_t - d_t$, the total amount of inventory possessed through the interval $[t, t+\tau]$. Note that $s_{t+\tau} = y_t - a_t - \sum_{j=0}^{\tau-1} d_{t+j}$. If we let $u_t = y_t - a_t$ be the *inventory position* through interval $[t, t+\tau]$ before ordering, then we have that $u_{t+1} = y_t - d_{t+1}$, which is familiar as the inventory difference equation under backlogging. Finally, if project out purchase costs the resulting

optimality equation is $v = \tilde{H}v$, where $\tilde{H}$ is defined to be:

$$\begin{aligned}\tilde{H}v(u,x) = \min_{y \geq u} \Bigg\{ & \mathbb{E}\left[\tilde{h}_\tau \left(y - \sum_{j=1}^{\tau} d_j\right)^+ + \tilde{p}_\tau \left(\sum_{j=1}^{\tau} d_j - y\right)^+ |x\right] \\ & + \beta \sum_{d',z'} \sigma(d',z'|x) v\big(y - d', \lambda(d',z',x)\big)\Bigg\},\end{aligned} \tag{3.4}$$

where $\tilde{h}_\tau = \beta^\tau h_\tau + c_\tau$ and $\tilde{p}_\tau = \beta^\tau p_\tau - c_\tau$. We detail this reformulation in Appendix B.3 and B.4.

3.4 Policy Construction

3.4.1 Base Stock Policies

Base stock policies have much appeal in the inventory literature due to their nice, simple structure that yields computational advantages, as well as easy practical implementation. We note that Equation 3.4 is in a familiar form, with backlogging state dynamics $u_{t+1} = u_t + a_t - d_t$ around the τ-lookahead inventory position (the total inventory orders across τ decision epochs into the future).

Define $y^*_\tau(x)$ as the smallest (and hence unique) myopic minimizer such that

$$y^*_\tau(x) \in \arg\min_y \left\{ \mathbb{E}\left[\tilde{h}_\tau \left(y - \sum_{j=1}^{\tau} d_j\right)^+ + \tilde{p}_\tau \left(\sum_{j=1}^{\tau} d_j - y\right)^+ |x\right]\right\}. \tag{3.5}$$

It will be convenient to define the inner function, g_τ:

$$g_\tau(y, d_1, \ldots, d_\tau) \triangleq \tilde{h}_\tau \left(y - \sum_{j=1}^{\tau} d_j\right)^+ + \tilde{p}_\tau \left(\sum_{j=1}^{\tau} d_j - y\right)^+.$$

In the following result, we provide conditions under which the base stock policy that orders $\max\{y^*_\tau(x) - u, 0\}$ when the current inventory position and belief vector are u and x, respectively, is optimal.

Proposition 19. *Suppose $y^*_\tau(x) - d' \leq y^*_\tau\big(\lambda(d',z',x)\big)$ for all d', z', x. Then the τ-lookahead*

policy, $\pi(u, x) = \max\{y^*_\tau(x) - u, 0\}$ *for all* u, x *is optimal.*

We note that the base stock levels are parametrized by the belief vector, x, and so thus far we still have considerable computational intractabilities, since the belief simplex X is uncountably infinite in cardinality. The size of X is a classic limitation of the POMDP framework. Additionally, we are left with the task of computing exactly the expectation within Equation 3.5 for each of these x, which may get unwieldy depending on the problem size. We discuss each of these issues and a method for addressing them, in turn.

The condition in Proposition 19 is an *attainability condition*. This condition guarantees that once $y^*_\tau(x_t) - u_t$ becomes non-negative, it will remain non-negative at all future epochs. An attainability condition first appeared in [56] and [57], was extended to the $\tau = 1$ case in [31], and is implicitly described in [54] as guaranteeing that the probability of excess inventory is zero.

Computing $\mathbb{E}[g_\tau(y, d_1, \ldots, d_\tau)|x]$**.** Now, we give a closed form for computing the expectation within Equation 3.5. Let $\mathcal{D}(d'|\mu', \mu) \triangleq \sum_{z'} P[z', d'|\mu', \mu]$. and $\mathcal{M}(\mu'|\mu) \triangleq P[\mu'|\mu]$. We now compute $P[d_1, \ldots, d_\tau|x]$:

$$\begin{aligned} P[d_\tau, \ldots, d_1|x] &= \sum_\mu x(\mu) P[d_\tau, \ldots, d_1|d, \mu] \\ &= \sum_\mu x(\mu) \sum_{\mu_\tau, \ldots, \mu_1} P[d_\tau, \ldots, d_1, \mu_\tau, \ldots, \mu_1|\mu] \\ &= \sum_\mu x(\mu) \sum_{\mu_\tau, \ldots, \mu_1} \mathcal{D}(d_\tau|\mu_\tau, \mu_{\tau-1}) \mathcal{M}(\mu_\tau|\mu_{\tau-1}) \mathcal{D}(d_{\tau-1}|\mu_{\tau-1}, \mu_{\tau-2}) \cdots \\ &\cdots \mathcal{M}(\mu_2|\mu_1) \mathcal{D}(d_1|\mu_1, \mu) \mathcal{M}(\mu_1|\mu), \end{aligned} \tag{3.6}$$

where the last equality follows from repeated application of the Markov property. Note that $\mathbb{E}[g_\tau(y, d, d_1, \ldots, d_\tau)|x] = \sum_{d_\tau, \ldots, d_1} P[d_\tau, \ldots, d_1|x] g_\tau(y, d_1, \ldots, d_\tau)$. This expression is a closed form representation of $\mathbb{E}[g_\tau(y, d, d_1, \ldots, d_\tau)|x]$ because the support of the demand distributions, the cardinality of the modulation space, and the lead time τ are all finite.

Even though we have a closed form for computing $\mathbb{E}[g_\tau(y, d, d_1, \ldots, d_\tau)|x]$, it may

be difficult to compute depending on the problem size. We note that each probability $P[d_1, \ldots, d_\tau | x]$ requires $|M|^\tau$ computations for every $(d_1, \ldots, d_\tau, x)$. Suppose the support of $\mathcal{M}[d|\mu', \mu]$ is Δ for all μ', μ. Then, for each x computing this expectation requires $\mathcal{O}(|M|^\tau \cdot |\Delta|^\tau)$ computations. Moreover, there are uncountably infinite belief vectors x. These computational considerations motivate a structural analysis of X and approximation methods for constructing a (nearly) optimal base stock policy.

Computing the base stock levels. As discussed earlier, one of the primary computational difficulties with any POMDP model is the size of the belief simplex, X. There are various methods in the literature that seek to address this issue, which may be broadly characterized as exact methods, fixed-grid approximation methods, and belief trajectory simulation methods. We focus our attention in this paper on finite grid approximations to X. The first, and most well-known, grid-based approximation procedure is found in [30], which builds an evenly-spaced finite grid approximation using Freudenthal triangulation controlled by a mesh parameter. In [22], a random finite grid approximation to X generated by points drawn from a uniform distribution on X is shown to perform comparably to the Freudenthal triangulation scheme. [5] introduced a trajectory following method in the context of passive learning in M-POMDPs that simulates the belief evolution of $\{x_t\}$ multiple times and constructs a finite grid based upon the most frequently simulated regions of X. Each of these grid-based procedures is applicable in our problem setting in a procedure as in Figure 3.1.

Direct application of these fixed grid procedures is independent of the structure of the base stock levels with respect to the belief vector, x. However, knowing this structure can be computationally useful. Generalizing results in [31] and assuming the demand distribution $\mathcal{M}[d|\mu', \mu]$ has finite support for all μ', μ, there is a finite partition of the belief space X such that the base stock level is identical for all x in each element of this partition and that each element is described by two linear equations in x, which we now show.

Let $\Delta_\tau \triangleq \{\sum_{j=1}^{\tau} \delta_j : \delta_1, \ldots, \delta_\tau \in \Delta\}$, the set of possible total demands over τ epochs.

1. Generate a finite grid approximation, $X_{finite} \subset X$ via any of the finite grid methods [30], [22], or [5].

2. For all points $x \in X_{finite}$, compute the optimal base stock level $y^*_\tau(x)$ by computing $P[d_1, \ldots, d_\tau | x]$ exactly or by estimating.

3. Define the base stock levels for the other points in $X \setminus X_{finite}$ by comparing to the computed nearest belief in X_{finite} (in the sense of some norm $\|\cdot\|$), so that $y^*_\tau(x) \approx y^*_\tau(x'(x))$, where $x'(x) \triangleq \underset{x' \in X_{finite}}{\arg\min} \|x - x'\|$.

Figure 3.1: Direct grid-based approximate base stock policy.

With some abuse of notation, let $\Delta_\tau \triangleq \{\delta_1, \ldots, \delta_{|\Delta_\tau|}\}$, where the δ_i are in ascending order ($\delta_1 < \delta_2 < \ldots < \delta_{|\Delta_\tau|}$). Additionally, let $\mathscr{D}_\tau$ be the τ-fold Cartesian product of Δ, so that elements of $\mathscr{D}_\tau$ are possible demand sequences $(d_1, \ldots, d_\tau)$. Let $\mathscr{D}_\tau(\delta)$ be the set of possible τ demand realizations that sum to less than or equal to δ:

$$\mathscr{D}_\tau(\delta) = \left\{ (d_1, \ldots, d_\tau) \in \mathscr{D}_\tau : \sum_{j=1}^{\tau} d_j \leq \delta \right\}, \quad \forall \delta \in \Delta_\tau.$$

We may use these sets $\{\mathscr{D}_\tau(\delta) : \delta \in \Delta_\tau\}$ as probabilistic events that allow us to generate a partition of X into sets X_m that are defined by the Newsvendor critical fractile, $\frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau}$:

$$\begin{aligned} X_m &\triangleq \left\{ x \in X : \sum_{\substack{(d_1,\ldots,d_\tau) \in \\ \mathscr{D}_\tau(\delta_{m-1})}} P[d_1, \ldots, d_\tau | x] < \frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau} \leq \sum_{\substack{(d_1,\ldots,d_\tau) \in \\ \mathscr{D}_\tau(\delta_m)}} P[d_1, \ldots, d_\tau | x] \right\} \\ &= \left\{ x \in X : P\left[\sum_{j=1}^{\tau} d_j \leq \delta_{m-1} | x \right] < \frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau} \leq P\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x \right] \right\}. \end{aligned}$$

The following proposition demonstrates that these regions X_m define the base stock levels, that is for every belief vector in X_m it is optimal to order-up-to δ_m.

Proposition 20. *The optimal base stock levels are $y^*_\tau(x) = \delta_m$ for all $x \in X_m$.*

Note that the partition of X is defined by hyperplanes since it is *linear* in x. Thus, in order to fully specify the base stock levels $\{y^*_\tau(x) : x \in X\}$, we need only construct at most

$|\Delta_\tau| - 1$ defining hyperplanes of the partition $\{X_m\}$. Moreover, these hyperplanes only need to be constructed once, and can be done *a priori* to any policy evaluation step.

These hyperplanes get more difficult to compute as τ gets larger. Thus, by the same reasoning as our earlier discussion about computing $P[d_1, \ldots, d_\tau | x]$, we seek an approximate method for generating the partition $\{X_m\}$ that scales well with τ. Our proposed method in Figure 3.2 is based upon finite grid approximations, Monte Carlo simulation, and solving $|\Delta| - 1$ support vector machines to determine the defining hyperplanes.

---

1. Generate a finite set of belief points, $X_{finite} \subset X$. Let $X_{finite} = \{x_1, \ldots, x_K\}$.
2. For each $x \in X_{finite}$, generate N demand trajectories $(d_t^n, \ldots, d_\tau^n)$. This gives us the estimate
$$\hat{P}\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right] = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\left\{\sum_{j=1}^{\tau} d_j^n \leq \delta_m\right\}.$$
3. Generate a label $l(x, m)$ for each x and m:
$$l(x,m) = \begin{cases} -1, & \hat{P}\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right] > \frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau} \\ 1, & \text{else.} \end{cases}$$
4. For each m, generate a separating hyperplane via a linear, soft-margin SVM on the set of tuples $\{(x_i, l(x_i, m)) : i = 1, \ldots, K\}$. The SVM is formalized by solving the following quadratic optimization problem, with penalty term C.
$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}\xi_i & \\ \text{s.t.} \quad & l(x_i, m)(\mathbf{w} \cdot x_i + b) \geq 1 - \xi_i, & i = 1, \ldots, K \\ & \xi_i \geq 0, & i = 1, \ldots, K \end{aligned}$$

Figure 3.2: Partitioning the belief space, X.

In Step 1, we generate a finite grid approximation $X_{finite} \subset X$ by either the Freudenthal triangulation of [30], the random grid of [22], or the trajectory following of [5]. In Step 2, we generate an estimate $\hat{P}\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right] \approx P\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right]$ by simulating demand trajectories, for each point in X_{finite}. Since we estimate these probabilities using Monte

Carlo methods, the computational burden scales linearly in τ, rather than exponentially when we compute the probabilities exactly.

We then label these points in Step 3, for each m according to which side $\hat{P}\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right]$ is on with respect to the critical fractile $\frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau}$. The regions are constructed in Step 4 using support vector machines (SVMs). For each m, a soft margin (due to the simulation error present in the estimates $\hat{P}[d_1, \ldots, d_\tau | x]$) linear SVM is computed, resulting in a hyperplane $\mathbf{w}^m \cdot x + b^m = 0$ that is defined by $(\mathbf{w}^m, b^m)$.

This hyperplane serves as a classifier for determining on which side of the critical fractile $\frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau}$ the probability $P\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right]$ falls. If $\operatorname{sgn}(\mathbf{w}^m \cdot x + b^m)$ is positive, then we predict that $P\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right] < \frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau}$. Likewise if $\operatorname{sgn}(\mathbf{w}^m \cdot x + b^m)$ is negative, we predict that $P\left[\sum_{j=1}^{\tau} d_j \leq \delta_m | x\right] \geq \frac{\tilde{p}_\tau}{\tilde{p}_\tau + \tilde{h}_\tau}$. By proceeding in this manner for all m, we determine our approximate order-up-to level for any x to be $y_\tau^*(x) \approx \delta_m$ such that $\operatorname{sgn}(\mathbf{w}^{m-1} \cdot x + b^{m-1}) \neq \operatorname{sgn}(\mathbf{w}^m \cdot x + b^m)$.

Moreover, we approximate the partition regions X_m to be

$$\begin{aligned} X_m &\approx \left\{x \in X : \operatorname{sgn}(\mathbf{w}^{m-1} \cdot x + b^{m-1}) \neq \operatorname{sgn}(\mathbf{w}^m \cdot x + b^m)\right\} \\ &= \left\{x \in X : \mathbf{w}^{m-1} \cdot x + b^{m-1} > 0, \mathbf{w}^m \cdot x + b^m \leq 0\right\} \end{aligned} \tag{3.7}$$

Thus, each region X_m is defined by two linear hyperplanes. We note that it is possible that the set $X_m = \varnothing$, which would indicate that δ_m is not an optimal order-up-to level for any $x \in X$. Thus, we have at most $|\Delta_\tau| - 1$ different non-empty regions of X.

3.4.2 Example Partition

In this subsection, we illustrate our method of generating the partition $\{X_m\}$ by considering a small example and analyzing the output. In this example, the modulation space has three elements $M = \{\mu_1, \mu_2, \mu_3\}$, the observation space has three elements $Z = \{z_1, z_2, z_3\}$, and the demand space has five elements $\Delta = \{1, 2, 3, 4, 5\}$. The dynamics $P[d', z', \mu' | \mu]$ are

governed by three matrices $\mathcal{M}$, $\mathcal{Q}$, and $\mathcal{D}$:

$$\mathcal{M} = \begin{bmatrix} 0.75 & 0.125 & 0.125 \\ 0.125 & 0.75 & 0.125 \\ 0.125 & 0.125 & 0.75 \end{bmatrix}, \qquad \mathcal{D} = \begin{bmatrix} 0.75 & 0.1 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.075 & 0.75 & 0.075 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.1 & 0.75 \end{bmatrix},$$

$$\mathcal{Q} = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix},$$

where $\mathcal{M}(i,j) = P[\mu' = \mu_j | \mu = \mu_i]$, $\mathcal{Q}(i,k) = P[z' = z_k | \mu = \mu_k]$, $\mathcal{D}(i,l) = P[d' = l | \mu = \mu_i]$, and $P[d' = l, z' = z_k, \mu' = \mu_j | \mu = \mu_i] = \mathcal{M}(i,j)\mathcal{Q}(i,k)\mathcal{D}(i,l)$. The lead time is $\tau = 2$, the discount factor $\beta = 0.9$, $\tilde{p}_\tau = 70$, $\tilde{h}_\tau = 10$ (which we denote as p, h for the remainder of this section).

Since $\tau = 2$, the set of possible total demand across τ is $\Delta_\tau = \{2, 3, \ldots, 10\}$ and, as in Section 3.4.1, the partitioning regions $\{X_m, m = 2, 3, \ldots, 10\}$ (we assume the indices m coincide with δ_m) are defined as:

$$\begin{aligned} X_m &= \left\{ x \in X : P[d_1 + d_2 \leq \delta_{m-1} | x] < \frac{p}{p+h} \leq P[d_1 + d_2 \leq \delta_m | x] \right\} \\ &= \left\{ x \in X : \Theta_{m-1} \cdot x < \frac{p}{p+h} \leq \Theta_m \cdot x \right\}, \end{aligned}$$

where we define $\Theta_m = \{\Theta_m(\mu_1), \Theta_m(\mu_2), \Theta_m(\mu_3)\}$ to be the vector such that $\Theta_m(\mu) = P[d_1 + d_2 \leq \delta_m | \mu]$. Thus, the true partitioning hyperplanes of X are of the form $\Theta_m \cdot x - \frac{p}{p+h} = 0$.

We use the procedure from Figure 3.2 to generate estimates of the partitioning hyperplanes of X using soft-margin, linear SVMs. The grid approximation of $X_{finite} \subset X$ is generated by randomly sampling 100 points from a uniform distribution on X and the estimates $\hat{P}[d_1 + d_2 | x]$ are generated by $N = 1000$ Monte Carlo simulations for each $x \in X_{finite}$. We generated the approximate partitioning hyperplanes of the form $\mathbf{w}^m \cdot x + b^m = 0$ using

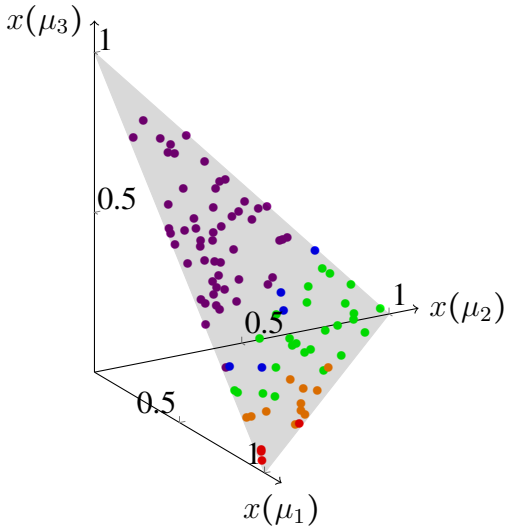
SVMs with penalty parameter $C = 10$ and $C = 50$. The results that define these approximate partitioning hyperplanes, and the true partitioning hyperplanes, are presented in Table 3.1. The entries for $\delta = 2, 3, 4, 5$ in Table 3.1 represent that there were no belief vectors $x \in X_{finite}$ such that $\hat{P}[d_1 + d_2 \leq \delta | x] > \frac{p}{p+h} = 0.875$. Thus, the labels of all the points in X_{finite} were the same (= 1). This is reasonable when we compare to the Θ vectors. We see for $\delta = 2, 3, 4, 5$, no entry of Θ is greater than the critical fractile $\frac{p}{p+h} = 0.875$; thus, the concomitant set X_m is empty. The implication is that it is never optimal to ever order-up-to a value less than or equal to 5.

Table 3.1: The SVM partitioning hyperplanes defined by $\{(\mathbf{w}, b)\}$ and the true partitioning hyperplanes defined by Θ and the critical fractile $\frac{p}{p+h} = 0.875$.

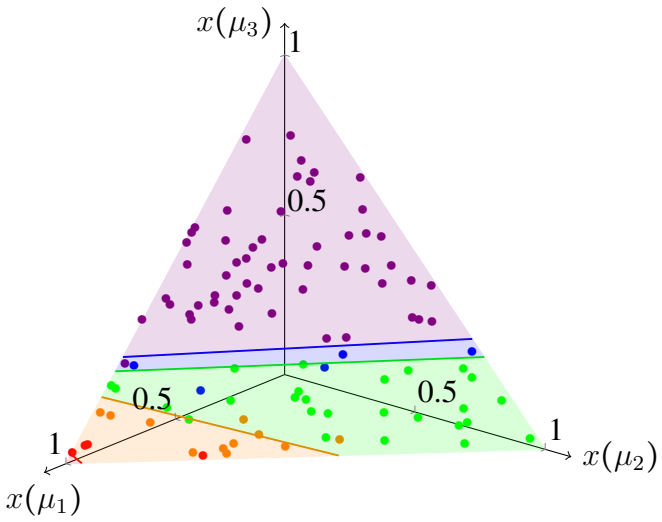
	$C = 10$		$C = 50$		
δ	$\mathbf{w}$	b	$\mathbf{w}$	b	Θ
2	–	–	–	–	$(0.43, 0.01, 0.01)$
3	–	–	–	–	$(0.56, 0.02, 0.02)$
4	–	–	–	–	$(0.70, 0.16, 0.03)$
5	–	–	–	–	$(0.79, 0.27, 0.06)$
6	$(-3.35, 0.77, 2.58)$	3.21	$(-7.56, -0.56, 8.12)$	6.23	$(0.94, 0.73, 0.21)$
7	$(-5.48, -1.82, 7.3)$	3.40	$(-9.58, -3.86, 13.45)$	5.82	$(0.97, 0.84, 0.30)$
8	$(3.59, 3.71, -7.30)$	−1.13	$(4.94, 6.70, -11.64)$	−2.02	$(0.98, 0.98, 0.44)$
9	$(4.01, 4.35, -8.35)$	−0.79	$(6.26, 6.46, -12.71)$	−1.23	$(0.99, 0.99, 0.57)$
10	–	–	–	–	$(1, 1, 1)$

The results of this partitioning method are depicted in Figure 3.3. Figure 3.3b depicts the SVM partition with $C = 10$, Figure 3.3c depicts the SVM partition with $C = 50$, and Figure 3.3d depicts the true partition of X. We color the points (belief distributions in X_{finite}) based on the estimated optimal order-up-to point (*i.e.* the m such that $\hat{P}[d_1 + d_2 \leq m - 1 | x] < \frac{p}{p+h} \leq \hat{P}[d_1 + d_2 \leq m | x]$ holds). Note that a color disparity between a point and the region to which it belongs is, in the case of the SVM partitions, either due to simulation error in the estimate of $\hat{P}[d_1 + d_2 \leq \delta | x]$ or an approximation error due to the dispersion and number of points in X_{finite}. In the case of the true partition of X, we know that these color mis-matches are due to simulation error in the probability estimates.

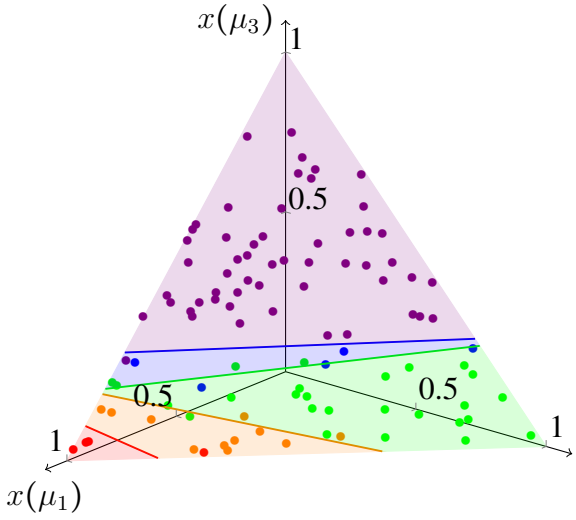
We note that as a rule of thumb, the SVM soft penalty term C should increase as the



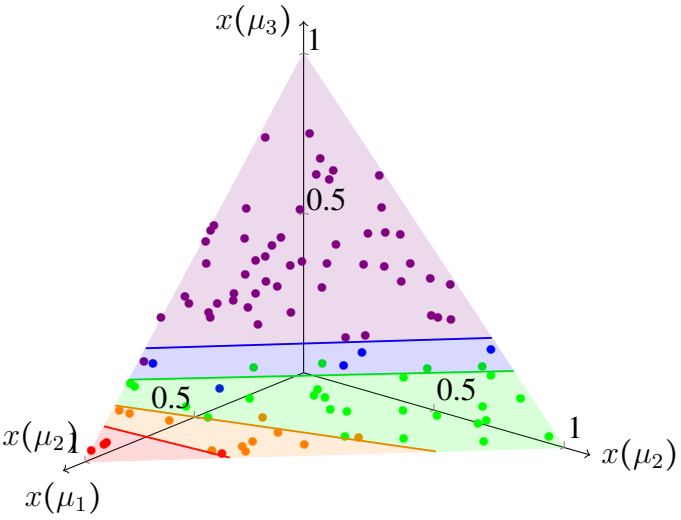
(a) Points randomly generated uniformly on the belief simplex, X, and labeled.



(b) The SVM-generated partition of X with C = 10.



(c) The SVM-generated partition of X with C = 50.



(d) The true partition of X.

Figure 3.3: Depicting the SVM partitions of X. The red regions correspond to δ = 6, the orange regions correspond to δ = 7, the green regions correspond to δ = 8, the blue regions correspond to δ = 9, and the violet regions correspond to δ = 10.

number of Monte Carlo simulation for estimating $\hat{P}[d_1, \ldots, d_\tau | x]$ increases since the estimate converges to the true probability almost surely for all $(d_1, \ldots, d_\tau)$. As C gets larger, the SVM discourages mis-classifications of the points $x \in X_{finite}$. Moreover, the SVM partition will improve as the number of grid points in X_{finite} increases.

Additionally, in this example we see that for a fixed $x(\mu_2)$, as the probability shifts from $x(\mu_1)$ to $x(\mu_3)$, the order-up-to level increases. This reflects two attributes of this example: (1) the construction of the $\mathcal{D}$ matrix, for which $\mathcal{D}[3, \cdot]$ has most of its probability mass on high demand levels and $\mathcal{D}[1, \cdot]$ has most of its probability mass on low demand levels, and (2) $\mathcal{M}$ has high probability mass on the diagonal. The implications are that if $x(\mu_3)$ is high, the DM believes there is a high probability of persisting in a high demand state into the future. Likewise if $x(\mu_1)$ is high, the DM believes there is a high probability of persisting in a low demand state into the future.

3.5 Application

In this section, we consider how the structural and algorithmic inventory policy results, above, might be used to facilitate an investigation of the following motivating question: should capital be allocated towards (1) a better information infrastructure (data, forecasts, quantitative talent, etc.), or (2) a more agile product architecture and supply chain design in order to more quickly respond to changing demand?

3.5.1 Value of Information.

It is intuitive that better information will improve supply chain performance. We seek to formalize this intuition by analyzing the *value of information* — namely, the relationship between the optimal value function and information quality. Before analyzing the value of information, we must conceptualize how this information is generated.

In Figure 3.4, we depict a *noisy channel* representation of the information process. At time t, the market — encompassing relevant macroeconomic factors pertaining to the firm — generates a hypothetical signal about its true, underlying state. This true signal undergoes a process of *data generation* that potentially corrupts the true market signal. Corruptions from data generation may come in the form of imperfect knowledge of the market, measurement errors, information loss due to data transmittal or summary, malevolent in-

terference, limited access, *etc*. This data then undergoes a *data processing* step — how the data is synthesized into information upon which decisions are made. This step includes forecasting, business intelligence insights, reporting, *etc*. We consider the composition of data generation and processing to be a noisy channel by which the true signal of the market is corrupted and AOD observations are generated. We call such an AOD observation mechanism the *information infrastructure*.

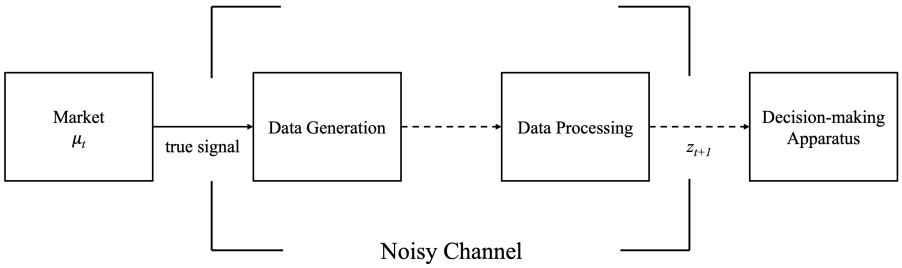


Figure 3.4: Noisy channel representation of the information infrastructure. Dashed lines represent distortions of the true signal.

In comparing two such information infrastructures (say q and $\tilde{q}$), we say that $\tilde{q}$ has *improved AOD observation quality* if it is equivalent to q when the observation generated by q is passed through an additional noisy channel. In other words, if AOD observations from the past information infrastructure are equivalent to corrupted AOD observations from the current information infrastructure, then the current infrastructure is an improvement over the past. This is consistent with the notion of improved observation quality, as presented in [45], [62], and elsewhere.

Definition 7. *For any two probability distributions $\tilde{q}(\tilde{z}|\mu', \mu), q(z'|\mu', \mu)$ over Z, we say that $\tilde{q}$ has improved AOD observation quality over q if there exists a $Z \times Z$ stochastic matrix, ξ, such that $q(z'|\mu', \mu) = \sum_{\tilde{z}} \xi(z'|\tilde{z})\tilde{q}(\tilde{z}|\mu', \mu)$.*

For the following proposition, suppose the demand and AOD processes are conditionally independent, so that $P[d', z'|\mu', \mu] = P[d'|\mu', \mu] \cdot P[z'|\mu', \mu]$. Let $q(z'|\mu', \mu) = P[z'|\mu', \mu]$ and $\tilde{q}(\tilde{z}|\mu', \mu) = \tilde{P}[\tilde{z}|\mu', \mu]$ be two different AOD probability measures on Z.

Additionally, let $\tilde{H}_q$ be the operator $\tilde{H}$ under q, $\tilde{H}_{\tilde{q}}$ be the operator $\tilde{H}$ under $\tilde{q}$, and let v and $\tilde{v}$ be the fixed points of $\tilde{H}_q$ and $\tilde{H}_{\tilde{q}}$ respectively.

Proposition 21. *Suppose $\tilde{q}$ has improved AOD observation quality over q such that there exists a Markov noisy channel ξ (i.e. a stochastic matrix $\{\xi(z'|\tilde{z})\}$) such that $q(z'|\mu', \mu) = \sum_{\tilde{z}} \xi(z'|\tilde{z})\tilde{q}(\tilde{z}|\mu', \mu)$. Then, $\tilde{v} \leq v$.*

Proposition 21 shows that if operating under an optimal policy, better information quality improves supply chain performance in terms of total discounted costs. Likewise, degradations of information quality increase costs. This effect also *typically* holds for near-optimal policies, as we demonstrate in the numerical exemplar of this section, although this is not necessarily true in all cases, as results in the literature show ([62]). For the manager, this shows the importance of efficiency in operations. If we consider the policy to be the mechanism by which information is turned into operational decisions, a bad policy can take better information and generate worse performance. Thus, in order to leverage the value of improved information the firm must adapt and adopt operations, accordingly.

3.5.2 Value of Agility

We now consider how system performance is dependent upon lead time, τ. Intuitively, longer lead times allow demand uncertainty in the supply chain to propagate and should thus generate worse system performance than shorter lead times, a phenomenon that we call the *value of agility*. We say system A is more *agile* than system B if (and only if) the lead time for system A is less than the lead time for system B. In our next result, we formalize this intuition and prove conditions under which it holds.

Let v_τ^* and $v_{\tau'}^*$ be the optimal value functions of the τ- and τ'-lagged problems respectively. Additionally, for compactness of notation, let $d_{[1:\tau'-\tau]} \triangleq (d_1, \ldots, d_{\tau'-\tau})$ and $z_{[1:\tau'-\tau]} \triangleq (z_1, \ldots, z_{\tau'-\tau})$.

Proposition 22. *Suppose $\tau < \tau'$, $\tilde{h}_\tau \leq \beta^{\tau-\tau'}\tilde{h}_{\tau'}$, and $\tilde{p}_\tau \leq \beta^{\tau-\tau'}\tilde{p}_{\tau'}$. Then,*

$$v_\tau^*\left(u - \sum_{j=1}^{\tau'-\tau} d_j, \lambda^{\tau'-\tau}\big(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x\big)\right) \leq \beta^{\tau-\tau'} v_{\tau'}^*(u, x)$$

for all demand realizations $d_{[1:\tau'-\tau]}$ and observations $z_{[1:\tau'-\tau]}$, where $\lambda^{\tau'-\tau}$ is the posterior belief distribution given x and $\tau' - \tau$ Bayesian updates from observations $d_{[1:\tau'-\tau]}$ and $z_{[1:\tau'-\tau]}$.

To understand this proposition, some accounting is required. Recall in our reformulation around inventory position, u, (Equation 3.4) that costs are accounted for at the time of the decision by projecting forward in time. Thus the costs accrued are scaled by the cost of capital so that they are reflective of what the costs would be if they were accrued on delivery. This is why the proposition includes the scaling factor $\beta^{\tau-\tau'}$. In this way, the proposition states that reducing lead time from τ' to τ decreases total costs under an optimal policy, when properly scaling for costs at any time t so that they are accounted for as if realized at time $t + \tau$ instead of $t + \tau'$.

The conditions on the per-unit costs in Proposition 22 are required due to the dependence of the single-period cost function on τ and have a natural interpretation. If $1 - \beta$ represents the cost of capital, then these conditions may be interpreted that cost restructuring under shorter lead times must improve per-unit costs by more than just the time value of money. Additionally, we note that, as with the value of information, this result holds under *optimal* policies. We show that this result holds for a near-optimal policy, as well, in our numerical example later in this section. For the manager, if investment is allocated to a supply chain restructuring, but the operating policy does not adapt to the new supply chain structure and the implications of the redesign on per-unit costs are not properly considered, then the value of agility may not materialize.

Finally, since Proposition 22 is dependent upon realizations of the demand and AOD

processes, it is natural to define the *expected marginal value of agility*, $m(u, x, \tau, \tau')$:

$$m(u,x,\tau,\tau') = \beta^{\tau'-\tau}\mathbb{E}\left[v^*_\tau\left(u - \sum_{j=1}^{\tau'-\tau} d_j, \lambda^{\tau'-\tau}\big(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x\big)\right)|x\right] - v^*_{\tau'}(u,x).$$

Note by Proposition 22 that $m(u, x, \tau, \tau') \leq 0$.

3.5.3 Numerical Exemplar

In our numerical example we consider a simple Markov chain model of the macroeconomy. We assume that there are 3 different states of the economy, with each state of the economy corresponding to a conditional demand distribution $\mathcal{D}[d'|\mu]$ that is characterized by its distributional characteristics (more specifically, its first two moments) as "high mean, high variance", "medium mean, medium variance", or "low mean, low variance". We denote these modulation states by μ_i where $i \in \{L, M, H\}$ represents that the demand distribution has low, medium, or high mean/variance, respectively. These modulation states might be considered to be a representation of various stages in the business cycle, in which the number of decision epochs for which the macroeconomy remains in a state μ_i is distributed geometrically with parameter $\mathcal{M}(\mu_i|\mu_i)$ for $i \in \{L, M, H\}$. We consider the following class of $|M| \times |M|$ probability transition matrices $\{\mathcal{M}_{\theta_M}\}$ parametrized by $\theta_M \in [0, 1]$:

$$\mathcal{M}_{\theta_M} = \begin{bmatrix} \theta_M & \frac{1-\theta_M}{2} & \frac{1-\theta_M}{2} \\ \frac{1-\theta_M}{2} & \theta_M & \frac{1-\theta_M}{2} \\ \frac{1-\theta_M}{2} & \frac{1-\theta_M}{2} & \theta_M \end{bmatrix}$$

The parameter θ_M, thus represents the probability of a state of the economy remaining in that state in the subsequent epoch. In this model, the expected number of epochs that the economy will stay in the same state is $\frac{1}{1-\theta_M}$ for $\theta_M \in [0, 1)$ and infinite if $\theta_M = 1$. Thus, the higher θ_M, the more stable we expect the economy to be.

In this numerical example, we consider $d_{t+1}|\mu_t$ to be distributed Poisson(3) if $\mu_t = \mu_L$, Poisson(6) if $\mu_t = \mu_M$, and Poisson(9) if $\mu_t = \mu_H$, truncated and scaled so that the demand

is less than 30 to ensure finiteness of the demand support.

The AOD observations are generated according to the conditional probabilities $\mathcal{Q}(\mu', z') = P[z'|\mu']$, which we describe by a class of stochastic matrices $\{\mathcal{Q}_{\theta_q}\}$ parametrized by θ_q, that is described by the base AOD matrix $\mathcal{Q}$:

$$\mathcal{Q} = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}.$$

We assume $\mathcal{Q}_{\theta_q} \triangleq \mathcal{Q}^{\theta_q}$. The parameter θ_q is, thus, the number of right matrix multiplications of the base AOD matrix, $\mathcal{Q}$. When $\theta_q = 0$, $\mathcal{Q}_{\theta_q}$ is the identity matrix. As θ_q gets large the matrix $\mathcal{Q}_{\theta_q}$ approaches the uniform stochastic matrix:

$$\lim_{\theta_q \to \infty} \mathcal{Q}_{\theta_q} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

In light of the value of information results in Section 3.4.3, we may view $\mathcal{Q}$ as a Markov noisy channel, where θ_q represents the number of times a perfect signal (*i.e.* $z_{t+1} = \mu_{t+1}$) is passed through the Markov noisy channel $\mathcal{Q}$. As θ_q gets large, this signal becomes uninformative as $z_{t+1}|\mu_{t+1}$ becomes uniformly distributed. In this interpretation, $\mathcal{Q}_{\theta_q}$ is monotone in θ_q in the sense that $\mathcal{Q}_{\theta_q}$ has *improved* AOD observation quality over $\mathcal{Q}_{\theta'_q}$ for all $\theta_q < \theta'_q$.

When describing the quality of AOD information for this parametrized class of observations matrices, $\{\mathcal{Q}_{\theta_q}\}$, we will use the language "ρ-perfect information quality". By this we mean the diagonal entry of $\mathcal{Q}_{\theta_q}$, call it ζ, scaled by its proportion of the interval $\left[\frac{1}{|M|}, 1\right]$ which reflects the extremes of completely uninformative AOD information, *i.e.* z_{t+1} is

distributed uniformly over M ($\zeta = \frac{1}{|M|}$), and perfect AOD information ($\zeta = 1$). Thus,

$$\rho = \frac{|M|}{|M|-1}\left(\zeta - \frac{1}{|M|}\right),$$

which gives a measure of the *degree* to which the AOD observations perfectly observe the modulation process. For instance, if the diagonal entry of $\mathcal{Q}_{\theta_q=1}$ is 0.9, then $\mathcal{Q}_{\theta_q=1}$ is $\rho = \frac{3}{2}(0.9 - \frac{1}{3}) = 85\%$-perfect AOD information quality.

In this numerical example we assume $\beta = 0.93$, which is based upon a current estimate of the weighted average cost of capital across all U.S. markets (approximately 7%, according to [26]). The true per-unit penalty cost is $p = 3$ and the true per-unit holding cost is assumed to be $h = 1$, yielding a critical fractile $\frac{p}{p+h} = 0.75$. We assume, for simplicity, that these per-unit costs are invariant with τ, so that $\tilde{p}_\tau = p = 3$ and $\tilde{h}_\tau = h = 1$. For our sensitivity analysis, we evaluate various lead times $\tau \in \{1, 2, 3, 4, 5\}$, parameters $\theta_M \in \{0.25, 0.5, 0.75, 1\}$ and $\theta_q \in \{0, 1, 2, 3, 4, 5\}$. We generate policies based on possible choices of per-unit stock-out penalty cost parameters $\theta_p \in \{3, 4, 5, 6\}$, and evaluate them using Monte Carlo simulation based on the true stock-out penalty term, $p = 3$. The motivation and numerical results pertaining to θ_p and stock-outs are discussed in Appendix B.5. Each policy is specified with respect to an SVM-generated partition of X, as in Sections 3.4.1 and 3.4.2 and detailed in Figure 3.2. The finite grid approximation $X_{finite} \subset X$ is a 1000 point grid randomly generated from a uniform distribution on X. The approximate probabilities $\{\Theta_m\}$ are generated on the basis of 1000 simulated demand trajectories.

We note that since our problem setting has an infinite horizon, we cannot simulate infinitely long trajectories to evaluate our policies. One potential method for evaluating policies via Monte Carlo simulation is to reformulate the infinite horizon problem as with a random geometrically-distributed time horizon, as in paper 5 of [37]. This would give a problem formulation that has an equivalent optimal policy and total expected discounted costs. However, we wish to analyze not only the estimated total costs, but also the variance

of the distribution and other metrics such as stock-out rates and violations of the attainability condition in Proposition 19. Reformulation around a random horizon would skew such metrics. Thus, we seek to simply evaluate a sufficiently long finite horizon, T. The error (in terms of total expected discounted costs) of such a finite horizon approximation is $\mathcal{O}(\beta^T)$, so we choose $T = 65$ in our evaluations so that $\beta^T = (0.93)^{65} < 1\%$.

Finally, we evaluated via Monte Carlo simulation the various parameter choices for θ_p, θ_q, θ_M, and τ, which amounts to $4 \times 6 \times 4 \times 5 = 480$ combinations. For each combination, we simulate the SVM-generated policy as in Figure 3.5, beginning from our assumed initialization of the belief vector over the states of the economy $x_0 = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$, the inventory on hand $s_0 = 0$, the epoch 0 demand $d_0 = 0$, and the history of actions $a_{-1} = \ldots = a_{-\tau} = 0$.

3.5.4 Sensitivity Analysis

Let $\theta \triangleq (\theta_p, \theta_M, \theta_q, \tau)$ be a combination of the Monte Carlo evaluation parameters. For each θ, we generate $N_{sim} = 1000$ samples $\{v_\theta^n : n = 1, \ldots, 1000\}$ for which v_θ^n is the realized total discounted costs in simulation n, for the base stock policy generated by the SVM-partitioning method of Sections 3.4.1 and 3.4.2 under parameter vector θ. We detail the method for generating the samples $\{v_\theta^n : n = 1, \ldots, 1000\}$ in Figure 3.5.

For each parameter vector θ, we collect the following metrics: mean value, standard error, and attainability violations.

- The mean value, which we denote v_θ, is the maximum-likelihood estimate of the expected total discounted costs under π_θ, the SVM-generated policy:

$$v_\theta = \frac{1}{N_{sim}} \sum_{n=1}^{N_{sim}} v_\theta^n.$$

- The standard error, which we denote SE_θ, is the maximum-likelihood estimate of the

0. *Initialize.* Initialize with the parameter vector $\theta = (\theta_p, \theta_M, \theta_q, \tau)$.

1. *Generate the SVM partition of X.* The partition of X is dependent on θ_p, θ_M, and τ, and is generated using the SVM method of Figure 3.2. Let $\mathrm{SVM}(x)$ be the order-up-to level of x according to the SVM partition.

2. *Monte Carlo simulation.* For each Monte Carlo simulation n, generate v_θ^n as follows.

   (a) Initialize $s_0^n = 0$, $x_0^n = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$, $d_0^n = 0$, $a_{-1}^n = \ldots = a_{-\tau}^n = 0$, and $v_\theta^n = 0$. Sample μ_0 from the belief distribution x_0.

   (b) For $t = 0, \ldots, T$:

   - *Determine ordering decision and cost.*

$$
\begin{aligned}
&y_t^n \leftarrow \mathrm{SVM}(x_t^n)\\
&u_t^n \leftarrow s_t^n - \sum_{j=1}^{\tau} a_{t-j}^n\\
&a_t^n \leftarrow \left(y_t^n - u_t^n\right)^+\\
&v_\theta^n \leftarrow v_\theta^n + \beta^t\left[\tilde{h}_\tau\left(s_t^n + a_{t-\tau}^n - d_t^n\right)^+ + \tilde{p}_\tau\left(d_t^n - s_t^n - a_{t-\tau}^n\right)^+\right]
\end{aligned}
$$

   - *Transition, costs, and belief update.*

$$
\begin{aligned}
&s_{t+1}^n \leftarrow s_t^n + a_{t-\tau}^n - d_t^n\\
&d_{t+1}^n \sim \mathcal{D}[\mu_t^n, \cdot]\\
&\mu_{t+1}^n \sim \mathcal{M}_{\theta_M}[\mu_t^n, \cdot]\\
&z_{t+1}^n \sim \mathcal{Q}_{\theta_q}[\mu_{t+1}^n, \cdot]\\
&x_{t+1}^n \leftarrow \lambda(d_{t+1}^n, z_{t+1}^n, x_t^n)
\end{aligned}
$$

Figure 3.5: The SVM-Monte Carlo evaluation method.

standard deviation of the total discounted costs generated by π_θ:

$$\mathrm{SE}_\theta = \sqrt{\frac{1}{N_{sim} - 1}\sum_{n=1}^{N_{sim}}\left(v_\theta^n - v_\theta\right)^2}.$$

- We denote by ATN_θ the number of violations of the attainability condition (in Propo-

sition 19) in the Monte Carlo simulation due to policy π_θ:

$$\mathrm{ATN}_\theta = \sum_{n=1}^{N_{sim}} \sum_{t=0}^{T} \mathbf{1}\left\{y_t^n > u_t^n\right\}.$$

The first two metrics — mean value and standard error — are meant to describe different aspects of a DM's *risk profile*. A DM's objectives may be multiple and may include minimizing costs (mean value) and minimizing tail risk (standard error). We seek to quantify the effect of input parameters on these objectives. We measure attainability violations in order to check the conditions necessary for a myopic base stock policy to be optimal that are presented in Proposition 19. In Appendix B.5, we consider stock-outs as an additional aspect to the DM's risk profile. There we also discuss how the DM might generate stock-out robustness by hypothetically manipulating their cost structure in constructing a policy, and the relationship between this approach and a Lagrangian relaxation of a chance-constrained version of our inventory problem.

Since the number of combinations of input parameters θ is large, we use regression analysis to describe how the parameters v_θ, SE_θ, and ATN_θ are related and propose for our initial regressions the following log-linear form (where SE_θ and ATN_θ may be substituted for v_θ):

$$\begin{aligned}\log(v_\theta) = b_0 + \sum_{j=1}^{5} b_{\tau=j}\mathbf{1}\{\tau = j\} + \sum_{j=1}^{4} b_{\theta_M=0.25j}\mathbf{1}\{\theta_M = 0.25j\} \\ + \sum_{j=3}^{6} b_{\theta_p=j}\mathbf{1}\{\theta_p = j\} + \sum_{j=0}^{5} b_{\theta_q=j}\mathbf{1}\{\theta_q = j\} + \epsilon,\end{aligned} \tag{3.8}$$

where ϵ is the error term. In these regressions, the endogenous variable (our metrics v_θ, SE_θ, SO_θ, and ATN_θ) is regressed against the *categorical* variables of our various input parameters. This regression is useful for estimating non-constant elasticities of our metrics with respect to the input parameters, θ. Note the following relationship between the regression estimated mean values under parameter vectors θ and θ', with the estimated regression

coefficients and metrics denoted with a hat (we drop the hat notation later when it is clear from context that we are referencing the estimated regression coefficients):

$$\frac{\hat{v}_\theta}{\hat{v}_{\theta'}} - 1 = \exp\left(\hat{b}_\tau - \hat{b}_{\tau'} + \hat{b}_{\theta_M} - \hat{b}_{\theta'_M} + \hat{b}_{\theta_p} - \hat{b}_{\theta'_p} + \hat{b}_{\theta_q} - \hat{b}_{\theta_q}\right) - 1. \tag{3.9}$$

Thus, the estimated percentage change in mean value (or other metrics) by changing from θ' to θ is an exponential function of the regression coefficients, so our regression model isolates the effects of τ, θ_M, and θ_q on the metrics that compose the DM's risk profile and the attainability condition of Proposition 19. In order to avoid over-fitting our regression model, we determine the subset of regressors for each metric using step-wise regression with Akaike information criterion. Summary metrics of these step-wise regressions are documented in Table B.1. The magnitude of the t-statistic indicates the level of statistical significance. In this type of regression, for each component parameter in θ, one of the categorical variables must be used as a reference variable, and thus may be interpreted to have a regression coefficient equal to 0.

On Attainability Violations. We first seek to verify the degree to which the attainability condition in Proposition 19 holds, which would add to the justification for analyzing base stock policies. In our analysis, we found that the attainability violation rate was relatively robust, with the vast majority of parameter combinations resulting in attainability violations that were less than 15%, as is shown in Figure 3.6. In our step-wise regression, the only significant parameters for attainability were τ and θ_M. For this example, attainability violations are shown to increase as τ increases and decrease as θ_M is closer to 0 or 1. Recall that attainability violations are dependent upon the evolution of the belief vector and the optimal base stock levels with respect to expected total discounted costs. Thus, our regression analysis shows that attainability violations increases with the uncertainty regarding the future states of the economy, either by uncertainty propagation through time (τ) or by the nature of the macroeconomic transition structure (θ_M). We note that attainability violations

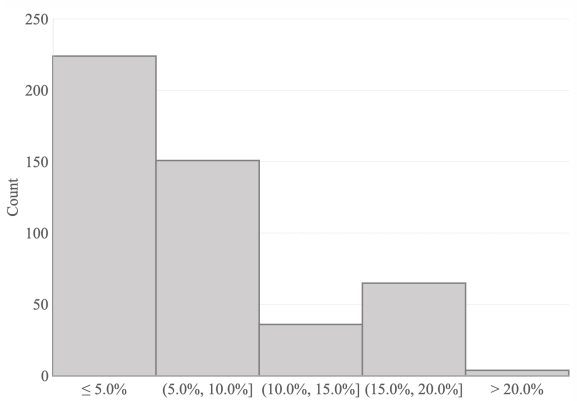


Figure 3.6: Histogram of attainability violations rates.

typically occur when the state of macroeconomy abruptly changes from a high mean demand state to a low mean demand state, leading to an oversupply of stock. Thus, the less an economic environment is subject to negative shocks, the better base stock policies will tend to perform.

On Expected Total Discounted Costs. The first, and primary, metric that we consider is the expected discounted costs. Recall that we discussed in Section 3.4.3 the relationship between expected total discounted costs and the AOD observation parameter, θ_q, as the *value of information*. We proved in Proposition 21 that improved AOD observation quality leads to lower expected total discounted costs under an optimal policy. Figure B.2a demonstrates the value of information. In this example, perfect AOD information — which corresponds to perfect knowledge of the state of the economy — is worth 2.4% in long-run cost savings over 85%-perfect AOD information quality (corresponding to $\theta_q = 1$), 3.9% in long-run cost savings over 72.25%-perfect AOD information quality (corresponding to $\theta_q = 2$), and 4.9% in long-run cost savings over 61.4%-perfect AOD information quality (corresponding to $\theta_q = 3$), and so on. Thus, better AOD information improves expected total discounts costs, as anticipated by Proposition 21. Further, in Figure B.2a note that v_θ appears concave in θ_q for the values we tested. This indicates that, at least for this numerical example, there is marginally increasing value in removing Markov noisy channels in AOD observations. This relationship holds when we additionally consider the predicted

value of AOD information with respect to our measure of ρ-perfect AOD information, as shown in Figure B.2b.

In Section 3.4.4, we discussed the relationship between lead time and expected total discounted costs as the *value of agility*. We prove that so long as the per-unit holding and penalty costs change with τ so as to provide cost savings by more than the time value of money, we expect long run costs to increase with τ. In Figure B.3b, this is exactly the effect observed. The undiscounted case represents the simulated example in which $\tilde{h}_\tau = 1$ and $\tilde{p}_\tau = 3$ for all τ. The discounted case represented these costs scaled by $\beta^{\tau-1}$ to represent an *estimate* of the costs under $\tilde{h}_\tau = \beta^{\tau-1}$ and $\tilde{p}_\tau = 3\beta^{\tau-1}$. The scale of long run cost savings with regard to reducing lead times is *much greater* in this example than improved information quality. For example, we estimate the value of reducing the lead time from $\tau = 5$ to $\tau = 4$ to be worth at least 17% in long run cost savings and reducing lead time from $\tau = 2$ to $\tau = 1$ is worth at least 36% in long run cost savings. We expect in practice that the price of such a restructuring of the supply chain or product design also scales similarly. Further, our estimated long run costs are convex in τ over the tested lead times, which suggests that, in absolute dollars, small reductions in long lead times leads to greater effects than small reductions in short lead times.

Finally, our numerical analysis demonstrates in Figure B.4 that, for this example, long run costs are lower as θ_M either approaches 0 or 1. This analysis reflects the intuition that the system should perform better in an economic environment that is more predictable. If θ_M is very small, then the DM will be confident that the economy will change in the next epoch. Likewise, as θ_M approaches 1, the DM will be confident in a stable, unchanging economic environment.

On Variance. For the DM concerned with either mitigating the upper tail risk of extreme scenarios that generate prohibitively high long run costs, we consider the sensitivity of the variance in long run costs to our various input parameters.

In Figure B.1, we summarize the regression analysis of the marginal effects of AOD

information quality on the standard error. Our regression analysis shows, for this numerical example, that better AOD information quality is effective for mitigating variance in long run costs. We quantify the magnitude of the marginal effects of AOD information quality on standard error in Figure B.1. However, this may be more simply observed in the histograms of sampled long run costs for various values of θ_q in Figure B.5. In comparing the relative peak heights of these unimodal distributions, we clearly note that these heights increase as θ_q gets smaller, which indicates that the variance of long run costs decreases with better AOD information quality. Moreover, this effect is amplified as AOD information quality improves, as depicted in the concavity of the normalized standard error curves with respect to θ_q and ρ in Figures B.1a and B.1b. Thus, small improvements to poor AOD information quality are less effective in mitigating tail risk than small improvements to good AOD information quality. Perfect AOD information quality, corresponding to $\theta_q = 0$ and $\rho = 1$, is estimated to reduce standard error of long run costs by 3.7% over 85%-perfect AOD information quality (corresponding to $\theta_q = 1$), by 5.0% over 72.25%-perfect AOD information quality (corresponding to $\theta_q = 2$), by 5.5% over 61.4%-perfect AOD information quality (corresponding to $\theta_q = 3$), *etc.*

As intuition might indicate, the most significant and determinative input parameter on variance is the lead time, τ. In Figure B.3, as we discussed in the relationship between τ and expected total discounted costs, the undiscounted case represents the simulated example in which $\tilde{h}_\tau = 1$ and $\tilde{p}_\tau = 3$ for all τ. The discounted case represented these standard errors scaled by $\beta^{\tau-1}$ to represent an *estimate* of the standard error under $\tilde{h}_\tau = \beta^{\tau-1}$ and $\tilde{p}_\tau = 3\beta^{\tau-1}$. Standard error is shown to decrease substantially as τ decreases. Moreover, similar to the result we found in the relationship between τ and long run costs, the marginal effect on standard error of increasing τ is amplified as τ increases. Reducing lead time from $\tau = 5$ to $\tau = 4$ is estimated to reduce standard error 22.2%; reducing lead time from $\tau = 2$ to $\tau = 1$ is estimated to reduce standard error 45.6%.

3.5.5 The Capital Allocation Process

In this subsection, we discuss how our numerical results may be used in a capital allocation optimization process. Consider the following business scenario as an example of how our results and methodology might be applied in a real-world context. Suppose a manager is tasked with the following objective: "first minimize costs and variance, while service level is above $1 - \alpha$, given a budget $\mathcal{B}$".

Let Θ be the set of scenarios $\theta = (\theta_p, \theta_q, \tau)$ initially considered by the manager (some of which may be infeasible *a posteriori*). For clarity, the parameters θ_p and τ are as defined in the numerical example of Section 3.5, and the parameter θ_q defines a particular AOD matrix $\mathcal{Q}_{\theta_q}$ that is not necessarily the same as any of the AOD matrices in Section 3.5. Additionally, let $\varphi(\theta)$ be the price of implementing scenario θ, relative to the current scenario θ_0. There are various ways the the manager might interpret the statement of the objective and use our methodology.

1. *Scalarize the objective.* The manager has access to a function $h : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $h(v_\theta, \mathrm{SE}_\theta)$ measures the relative value the manager places towards the two objectives of long run costs and variance, subject to constraints on stock-outs and budget feasibility. This method is often referred to in the multi-objective optimization literature as *scalarizing* the objective.

$$\begin{aligned} \min_{\theta \in \Theta} \quad & h(v_\theta, \mathrm{SE}_\theta) \\ \text{s.t.} \quad & \mathrm{SO}_\theta \leq \alpha \\ & \varphi(\theta) \leq \mathcal{B} \end{aligned}$$

2. *Prioritize one objective.* The manager chooses one of the objectives to be primary, say long run costs, and treats the second objective, say variance, as a constraint such

that the variance is improved relative to the current scenario.

$$\begin{aligned} \min_{\theta \in \Theta} \quad & v_\theta \\ \text{s.t.} \quad & \mathrm{SO}_\theta \leq \alpha \\ & \varphi(\theta) \leq \mathcal{B} \\ & \mathrm{SE}_\theta \leq \mathrm{SE}_{\theta_0} \end{aligned}$$

3. *Pareto optimality.* In this interpretation, the manager does not have access to a scalarizing function h, so instead solves for a set of scenarios that are *Pareto optimal*. Thus the problem is to solve for the Pareto optimal set, $\mathrm{PO}(\Theta)$:

$$\mathrm{PO}(\Theta) \triangleq \left\{ \theta \in \Theta : \mathrm{SO}_\theta \leq \alpha, \varphi(\theta) \leq \mathcal{B}, \{\theta' \in \Theta \setminus \theta : v_{\theta'} > v_\theta, \mathrm{SE}_{\theta'} > \mathrm{SE}_\theta\} = \varnothing \right\}$$

There may be additional feasibility constraints to the system, which if known may be included as explicit constraints. If unknown, but intuited by the manager, then the these constraints may be treated as *post hoc* cuts. For example, in the first two optimization approaches, the manager might proceed in an iterative manner, as in the familiar cutting plane methods in mathematical programming. The manager might solve the optimization problem, then determine whether the resulting solution is feasible, and if not add a cut, re-solve, and proceed until a solution that is feasible to the manager emerges. A benefit of the third approach is that the Pareto frontier, a set of solutions, is generated. The manager can then simply determine the subset of these Pareto optimal scenarios that are feasible to implement.

Each of these budget allocation optimization formulations has many different solution approaches, and it is outside the scope of this paper to consider how to solve these optimization problems. We formulate the budget allocation problem in different ways simply to note that each of these requires *as input* $(v_\theta, \mathrm{SE}_\theta, \mathrm{SO}_\theta)$ for all $\theta \in \Theta$. Thus, the manager

has at least two options for using our methodology:

1. *Enumerative Evaluation, then Optimization.* Evaluate $(v_\theta, \mathrm{SE}_\theta, \mathrm{SO}_\theta)$, as described in Figure 3.5, for all $\theta \in \Theta$. Then, solve an optimization problem that takes as input these evaluations.

2. *Representative Evaluation, Regression, then Optimization.* Evaluate $(v_\theta, \mathrm{SE}_\theta, \mathrm{SO}_\theta)$ for some subset of possible parameters $\Theta_{rep} \subset \Theta$. Utilize a regression akin to Equation 3.8 (or suitably modified) as a model for determining the estimates of $(v_\theta, \mathrm{SE}_\theta, \mathrm{SO}_\theta)$ for all $\theta \in \Theta \setminus \Theta_{rep}$. Then, solve an optimization problem that takes as input these evaluations and estimations.

The first method is a straightforward application of our evaluation methodology for all scenarios $\theta \in \Theta$, which is computationally tractable for small scenario sets Θ or time-insensitive application settings. For more time-sensitive applications or large scenario considerations, the second method uses a representative subset of scenarios and a regression to formulate approximate evaluations of $(v_\theta, \mathrm{SE}_\theta, \mathrm{SO}_\theta)$. This approach is related to value function approximation methods in the approximate dynamic programming literature ([36]).

Finally, we note that since we have demonstrated a method for quantifying the monetary effects of AOD information, lead times, and stock-out penalties, the manager may use our methodology to compute the return on investment (ROI) of each scenario, $\mathrm{ROI}(\theta)$:

$$\mathrm{ROI}(\theta) = \frac{v_\theta - v_{\theta_0}}{\varphi(\theta)}.$$

3.6 Future Research Directions

There are a number of future research directions for this work. In this paper, we model data as the end result of a noisy channel composed of data generation and processing. This allows us to generalize the concept of data to include anything from economic indicators to

demand forecasts and define explicitly a notion of the quality of this information in terms of composed noisy channels. This generalization gave us rich mathematical structure as well, which we used in Proposition 21 to establish the value of information. However, our notion of quality in this paper is relative (comparing one information infrastructure to another), and the concept of noisy channels as a measure of quality — though analytically and conceptually appealing — is a bit unintuitive to consider in practice. It's more natural to consider instead an absolute measure of quality, say on $[0, 1]$, so the practitioner can consider information to $x\%$ perfect. We introduce the ρ metric in Section 3.5 for our numerical example for this reason. Incidentally, for our numerical example, the ρ metric we introduced is (1) an eigenvalue of $\mathcal{Q}_{\theta_q}$ and (2) the absolute value of the correlation between μ_t and z_t. An interesting line of future research is to investigate how these or other measures might be generalizable, so that we might measure the quality of information absolutely.

Additionally, there are other operational problem settings for the data-driven stochastic dynamic program that may be of interest to pursue, as we discussed in Section 3.2. Finally, we note that mean value, standard error, stock-outs, and attainability are a sample of the possible metrics that might be considered in the Monte Carlo evaluation of Section 3.5. Depending on the business setting and the DM's risk profile, it may be of interest to consider alternative metrics in addition to, or in lieu of, the metrics we investigate.

CHAPTER 4

GENERATING TRUST IN DEVELOPMENT PROCESSES USING ROBUST, DATA-DRIVEN MARKOV GAMES: AN APPLICATION TO PRESTIGE

4.1 Introduction & Literature Review

4.1.1 Introduction

In many real-world applications, we are concerned with how raw goods, ideas, tasks, *etc.*, progress from an initial state and develop towards an end goal. Supply chains take raw materials — and through a network of steps including transportation, production and manufacturing, warehousing, assembly, *etc.* — generate an end product. Development chains take initial product concepts and materials and generate a new product design. Maintenance systems perform routine maintenance and/or take faulty or outdated equipment and undergo a restorative process of re-engineering, hardware/software updates, *etc.* to generate updated, newly functioning equipment. Many artificial intelligence (AI) systems likewise undergo a sequence of steps in order to learn and automate a given task. We call functioning systems such as these, in which a series of events must take place in order to reach an end goal (*i.e.* a product design, a product in the market, a refurbished piece of equipment, an AI system), *development processes*. In this paper, we are concerned with so-called *trust* problems. That is, we answer the following motivating question: how much can we *trust* the end result of a development process to function as it is properly intended?

The research in this paper is motivated by efforts to incorporate optimal or near-optimal dynamic decision making capability based on inaccurate observations into the PRESTIGE (PRactical Evaluation and Synthesis of Trust In Government systEms) decision support system developed at Sandia National Laboratories. The modeler or user (or *defender*, the term we use in our discussion of our game-based model formulation) of the PRESTIGE

decision support system faces a development process — a logical sequence of steps necessary to achieve an end goal — in which there is potential for adversarial manipulation, and must make decisions on the basis of possibly noise-corrupted data. In such situations, there is uncertainty or imprecision as to the objectives, resources, and level of rationality of a potential adversary. Moreover, whereas in many adversarial learning problems there is access to training samples, development processes may be one-off processes and the risk of adversarial manipulation a previously unknown or scarcely observed occurrence. In the presence of adversarial training samples, the risk of misapprehension of attacker objectives and resources may be mitigated by learning the attacker's policy through data. However, the lack of adversarial training samples in many development process applications makes the risk due to misapprehension of attacker objectives all the more acute, since the attacker's objectives will ultimately determine their policy. In such problems, we seek to generate a single defender policy that specifies at each decision epoch, on the basis of the information available, actions for achieving progress towards the end goal of the development process while ensuring that the output of the development process is as trustworthy as possible. Such a policy would be a *trusted* or *trustworthy* policy. More specifically, in defending a development process from potential adversarial manipulation, we seek a policy that (1) achieves progress towards the end goal of the development process, and (2) can deter attacks and if necessary, detect, respond to, and recover from an attack so that the output of the development process is as trustworthy as possible.

Thus, we recognize that generating trust in these development processes is inherently a multi-faceted concept. It is inextricably tied to risk mitigation — risk due to adversarial manipulation, misapprehension of adversarial objectives and resources, natural processes, model misspecification, *etc.* A trustworthy policy, then, must be *robust* — that is, the policy reliably accomplishes what it is intended to do, and is secure with respect to potential adversarial manipulation. Moreover, since the policy specifies how the defender behaves dynamically over time and is generated algorithmically, we seek a policy whose behavior

is inherently *explainable* to interpreters of the model. That is, in constructing the policy, does our algorithm confer information about why the policy proscribes certain actions? These notions of *robustness* and *explainability* are important concepts in developing trust in artificial intelligence systems ([21]), and we discuss the implications of our research on artificial intelligence throughout.

In this paper, we first show that the general *trust* problem can be modeled using a definition of state and state dynamics based on a development process graph and a generalized concept of an attack graph that we call a *precedence graph*. We then model the dynamic decision-making problem in which two known, intelligent, and adaptive agents — an attacker and a defender — interact for control of a development process using the partially-observable Markov game (POMG) framework ([9], [8], [10]). This serves as a model of how each agent would interact, *if* the attacker's objectives, resources, and level of rationality were known. Since this is not generally the case in *trust* problems, we propose a three-fold heuristic solution procedure:

1. We use a relaxation of the POMG model as a *generative mechanism* for creating *hypothetical* or *potential* attacker policies by considering differing attacker objectives and orders of rationality.

2. We use a robust dynamic program that explicitly incorporates these potential attacker policies in constructing a robust policy, under perfect defender information.

3. The defender plays a *probability matching* heuristic on the basis of the robust policy and the defender's belief about the state of the development process.

We then show how this robust heuristic policy may be evaluated using Monte Carlo simulation, and evaluate its performance on a numerical exemplar. Finally, we note that as reinforcement learning-based AI systems are foundationally dependent upon Markov decision processes ([4]), and seek to determine an optimal policy, we see potential for foundationally important advancements in generating trust in artificial intelligence systems through

explicit incorporation of robustness and explainability in policy generation, as we present in this paper. We see extending the dynamic programming (DP) methods in this paper to incorporate approximate DP and heuristic search methods as potentially fruitful research toward these aims. We discuss these potential contributions in the concluding section.

4.1.2 Literature Review

This research draws from ideas in various strands of literature.

PRESTIGE. PRESTIGE is a discrete event and Monte Carlo simulation-based tool set based on a model of attacker-defender interaction called GPLADD (Graph-based Probabilistic Learning Attacker and Dynamics Defender) described in [15] and [33]. GPLADD represents an attack via multiple stochastic games, called PLADD (Probabilistic Learning Attacker and Dynamics Defender, [24] and allows analysis of attack success parameters, the effects of the defender strategy, and defender strategy optimization. GPLADD has been applied to the study of data injection attacks on the electric power grid in [11] and elements of PLADD has been applied to the same problem in [12]. Effects of detection in GPLADD games are discussed in [18]. GPLADD games are constructed from individual PLADD games. In PLADD, both a single attacker and a single defender compete for a resource. Defender actions include the ability to retake control of the resource from the attacker and disrupt any knowledge acquisition accrued by the attacker. For both agents, all actions accrue cost, and rewards result from the length of time the resource is under the control of the agent. PRESTIGE models attacker actions with an attack graph. Attack graphs include actions to undermine access controls. In this paper, we consider the general problem setting of PRESTIGE and incorporate dynamic decision-making and Bayesian inference based upon noise-corrupted data.

POMG. In this research, we use the POMG as the fundamental model of dynamic, data-driven, and interactive decision-making between a defender and an attacker. The POMG was originally introduced in [8] and [9] as a generalization of the multi-period stochastic

game to partially observed decision-making, analogous to the way the partially observed Markov decision process (POMDP) is a generalization of the Markov decision process (MDP). The POMG was further applied to security in food production processes in [10]. We give a formal introduction of the POMG framework in Section 4.2.

Robust Optimization. In this paper, we are concerned with developing defender policies that are robust to a range potential attacker policies, in order to develop a singular defense policy that generates trust in the development process, regardless of the type of adversary. [51] introduced a generalization of the MDP, the *multi-model MDP* (MMDP), in which there is a set of candidate models of transition probability or cost function structures that may be based on multiple competing data sources. The decision-maker (DM) is tasked with determining a single policy that optimizes the expected weighted performance over all models. In this way, a solution to the MMDP is a policy that is robust to model uncertainty. In Sections 4.4 and 4.5, we show that each attacker policy induces a best response stochastic dynamic program (DP) for the defender. In Section 4.5.2, we show how we may adapt results in [51] to determine a defender policy that is robust to an array of attackers by solving a best response MMDP.

Training and Genetic Algorithms. In Section 4.5.1, we introduce an iterative method of training the agents in order to generate a set of plausible and adaptive attacker policies. At each iteration, a new generation of candidate attacker policies is generated by pairing the attacker with a defender policy and solving a best response MDP. We discuss how this method of training is similar to the selection and mutation mechanism of genetic algorithms, introduced in [23]. Moreover, the concept of training adversarial agents using successive competition is similar to and inspired by recent work in [44] using neural networks and competitive self-play to develop policies for playing Go, shogi, and chess.

Thompson Sampling. We develop a randomized heuristic policy for the defender with partial observability into the state of the system that has ties to Thompson sampling, a popular machine learning technique for multi-armed bandit problems that was first introduced

in [52] and has been adapted to many operations research applications ([39], [17]). In the appendix, we elucidate the connection between our heuristic and a non-stationary variant of Thompson sampling. For a more in-depth treatment of this topic we refer the reader to [40].

4.2 An Overview of the POMG

The POMG is a model of multi-agent sequential decision making under uncertainty and partial (inaccurate, noise corrupted) observations. In the context of the *trust* problems, there are two agents — an attacker and a defender. Each agent selects an action at each of a countable number of decision epochs, $t = 0, 1, \ldots, T$, where $T = \infty$ corresponds to an infinite horizon POMG and $T < \infty$ corresponds to a finite horizon POMG. An agent's actions are chosen with the intent of minimizing the agent's criterion and are based on the agent's current state of knowledge. An agent's current state of knowledge can contain past and present observations of the state of the system and these observations may be partial (inaccurate, noise corrupted). The dynamics and criterion of each agent are assumed to be affected by the actions of both agents. Thus, each agent must consider what actions the other agent might select in order to minimize its criterion.

More precisely, the POMG includes the following stochastic processes for each agent k (either the attacker $\mathcal{A}$ or the defender $\mathcal{D}$):

- $\{S_t^k\}$ represents the *state* of agent k at epoch t,
- $\{Z_t^k\}$ represents the *observations* that agent k receives about the system at epoch t, and
- $\{A_t^k\}$ represents the *actions* taken by agent k at epoch t.

We assume that agent k selects A_t^k on the basis of $\mathscr{I}_t^k$, the state of knowledge or *information pattern* of agent k at epoch t. This action is chosen from the set of admissible

actions $\mathscr{A}^k$ that may or may not depend on $\mathscr{I}_t^k$. We assume $\mathscr{I}_t^k$ contains Z_τ^k, the observation agent k receives about the other agent just before epoch $\tau < t$. The dynamics of the system are described by the given probabilities $P[Z_{t+1}, S_{t+1}|S_t, A_t]$, where $S_t = (S_t^{\mathcal{A}}, S_t^{\mathcal{D}})$, $A_t = (A_t^{\mathcal{A}}, A_t^{\mathcal{D}})$, and $Z_t = (Z_t^{\mathcal{A}}, Z_t^{\mathcal{D}})$. The state, action, and observation spaces for one agent may be totally different from the state, action, and observation spaces for the other agent; however, we assume all of these sets are finite.

For agent k, we assume the cost accrued between epochs t and $t+1$ is $C^k(S_t, A_t)$ and the criterion is the expected total discounted cost over a finite or infinite horizon. A policy π^k for agent k is a mapping from the set of all $\mathscr{I}_t^k$ to the set of all A_t^k, i.e., $A_t^k = \pi^k(\mathscr{I}_t^k)$, for all t. (We remark that a generalization of this definition is to let a policy be a mapping from the set of all $\mathscr{I}_t^k$ to the set of all probability distributions over the set of all A_t^k, a generalization that will prove useful when we develop heuristics in Section 4.5.) Thus, we seek a policy — a rule that tells the agent what action to take, given the agent's current state of knowledge. Let Π^k be the set of feasible policies for agent k. Given policy π_j, $j \neq k$ and $j, k \in \{\mathcal{A}, \mathcal{D}\}$, the criterion for agent k is to minimize the expected total discounted costs, $V_0^k(\mathscr{I}_0|\pi^j)$:

$$\begin{aligned} V_0^{\pi^k}(\mathscr{I}_0^k|\pi^j) &\triangleq \mathbb{E}\left[\sum_{t=0}^{T} \beta^t C^k\big(S_t, \pi^k(\mathscr{I}_t^k), \pi^j(\mathscr{I}_t^j)\big) + C_{T+1}^k(S_{T+1})|\mathscr{I}_0\right], \\ V_0^k(\mathscr{I}_0^k|\pi^j) &= \min_{\pi^k \in \Pi^k} V_0^{\pi^k}(\mathscr{I}_0^k|\pi^j). \end{aligned} \tag{4.1}$$

where $\beta \in [0, 1]$ is the discount factor. For the finite horizon setting when $T < \infty$, we take C_{T+1}^k to be the terminal cost function for agent k. For the infinite horizon setting when $T = \infty$, then β must be strictly less than 1, and we let $C_{T+1}^k = 0$.

We remark that the value of one agent's criterion is dependent on the policies of *both* agents. Thus, ideally a solution to this problem would be a Nash equilibrium, *i.e.* a policy pair $(\pi^{\mathcal{A}}, \pi^{\mathcal{D}})$ such that $V_0^{\mathcal{A}}(\mathscr{I}_0^{\mathcal{A}}|\pi^{\mathcal{D}}) = V_0^{\pi^{\mathcal{A}}}(\mathscr{I}_0^{\mathcal{A}}|\pi^{\mathcal{D}})$ and $V_0^{\mathcal{D}}(\mathscr{I}_0^{\mathcal{D}}|\pi^{\mathcal{A}}) = V_0^{\pi^{\mathcal{D}}}(\mathscr{I}_0^{\mathcal{D}}|\pi^{\mathcal{A}})$. A more detailed description and analysis of the POMG can be found in [9], [8], [10].

4.3 Modeling Trust problems with the POMG

Graphs can be useful in the development of well-defined POMGs for many *trust* and *trust*-related problems. For example, in modeling the food production security problem considered in [8] and [10], two distinctly different AND/OR graphs, one for the attacker and one for the defender, were useful in representing important aspects of the problem, and these graphs serve as an intermediate step in the development of a well-defined POMG.

For POMG model construction, many *trust* problems can be (at least partially) described by one or more directed acyclic OR graphs. These problems are modeled with a *development process* graph, which is associated with the defender, and a set of attack graphs, which is associated with the attacker. An attack graph describes the actions the attacker is to take and the order in which these actions are to be taken. In this paper, we generalize the notion of an attack graph to a *precedence* graph, an acyclic OR graph that describes all actions the attacker can take and the order in which these actions can be taken, *i.e.* the union of the attack graphs. The development process graph and the precedence graph share nodes, which we call the *game* nodes since they are the nodes in which attacker and defender interact with each other. Game nodes consist of *access control* nodes and *artifact* nodes. Artifact nodes correspond to objects that are contested, and access controls correspond to gateways that protect or grant access to these objects. For example:

- In a cyber-network context, an artifact might be RTL files and an access control might be an encrypted hard drive on which the RTL files are stored.
- In a bank, an artifact might be deposited currency and an access control might be a vault door.

Interactions may take place between the attacker and defender at the perimeter level of access controls, or at the level of artifacts themselves.

Let $N_{\mathcal{A}}$ and $N_{\mathcal{D}}$ be the nodes in the precedence and development process graphs, respectively. The node $n_{\mathcal{A}} \in N_{\mathcal{A}}$ is the start node for the attacker. Nodes with null successor

sets are called terminal nodes. Let $N' = N_{\mathcal{A}} \cap N_{\mathcal{D}}$ be the game nodes. The attacker's start node is neither a game node nor a terminal node. The intersection of the set of non-game nodes for the attacker and the set of non-game nodes for the defender is null.

Let $E_{\mathcal{A}} \subset N_{\mathcal{A}} \times N_{\mathcal{A}}$ and $E_{\mathcal{D}} \subset N_{\mathcal{D}} \times N_{\mathcal{D}}$ be the directed edges in the precedence and development process graphs, respectively. The meaning of an edge is domain dependent. For example, an edge (n, n') in the precedence or development process graph might indicate the existence of a situation having a non-zero probability of an agent successfully taking control of node n', given the agent has control of node n.

We call the union of the precedence and development process graphs the *unified graph*.

The State of the Unified Graph. The state of a graph is represented by the state of the nodes in the graph. We now describe the state of a game node and a non-game node.

A game node can be in one of three states:

- The access control or artifact does not yet exist (state 0)
- The access control or artifact exists and is under the control of the attacker (state $\mathcal{A}$)
- The access control or artifact exists and is under the control of the defender (state $\mathcal{D}$).

The access control or artifact is created when the defender first visits the game node, and the defender is immediately in control of the node. Thus, once the defender successfully reaches the game node for the first time, the node makes transition from state 0 to state $\mathcal{A}$. Once a game node's state leaves state 0, it will never return to state 0 but thereafter will be in one of the two states in the set $\{\mathcal{A}, \mathcal{D}\}$.

The state of the start node for the attacker is always $\mathcal{A}$. A non-game node in the precedence graph that has been visited by the attacker is in state $\mathcal{A}$ and otherwise is in state 0. Similarly, a non-game node in the development process graph that has been visited by the defender is in state $\mathcal{D}$ and otherwise is in state 0.

Let $s_t(n)$ be the state of node n at epoch t. Then, the state of the unified graph at epoch t is $S_t = \{s_t(n) : n \in N_{\mathcal{A}} \cup N_{\mathcal{D}}\}$. Initially (at epoch 0), all non-start nodes are in state 0 for

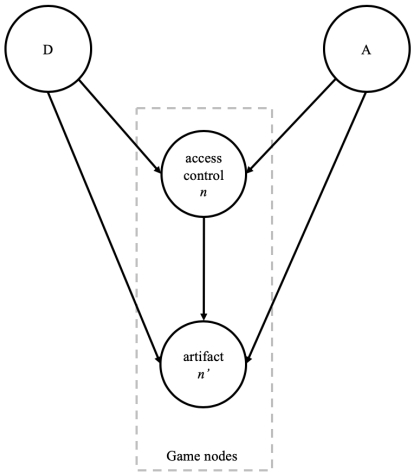


Figure 4.1: *Modeling the existence phenomena using context.* At initialization access control n and artifact n' are in state 0, *i.e.* they do not exist. We define the context of the attacker take move on n corresponding to the edge (A, n) to be simply that n is not in state 0. The context of the attacker take move on n' from A, *i.e.* a take move directly on the artifact n', corresponding to edge (A, n') is that n is in state 0 and n' is not. In other words, the defender created the artifact n' before creating the access control. Once n is not in state 0, *i.e.* it has been created by the defender, the direct take move on n from A is not permissible.

both graphs, which is an easily modified assumption.

Action Selection. A variety of actions can be available to each agent. We mention three possible actions:

- *Take.* A successful Take move on a node n by agent k at epoch t results in node n being under the control of agent k at epoch $t + 1$. Each agent can only attempt to control nodes in its own node set. We remark that a Take move in this paper is a slightly different notion than the Take move in GPLADD ([18]).

- *Protect.* Each agent can decide to reduce the probability of success of a Take move on a game node by its adversary (corresponds to a delay under a memoryless assumption) by the Protect action.

- *Observe.* Each agent may choose to observe the state of nodes in either node set. These observations may be noise corrupted.

Observations. We note that the Take and Protect actions affect the state transitions of the development process and precedence graphs. The Observe action affects the information available to the agents. Let $Z_t = \{z_t(n) : n \in N_{\mathcal{A}} \cup N_{\mathcal{D}}\}$, where $z_t(n)$ is the possibly noise corrupted observation of $s_t(n)$. We assume that $P[Z_{t+1}|S_{t+1}, S_t, A_t]$ is given, where A_t is the collection of actions taken by the two agents at epoch t.

The Transition Structure of the State of the Unified Graph. Take actions on the part of both agents can affect the state of both graphs in making state transitions, *i.e.*, in changing from $S_t^{\mathcal{A}}$ and $S_t^{\mathcal{D}}$ to $S_{t+1}^{\mathcal{A}}$ and $S_{t+1}^{\mathcal{D}}$. In order for an action to change the state of a specific node n to be successful, a necessary (but not sufficient) condition is that the current states of the *ancestor* nodes, of node n, which are defined to be all of the nodes n' in the precedence (development process) graph such that there exists a path from $n_{\mathcal{A}}$ ($n_{\mathcal{D}}$) to n' that does not include n, are in *permissible* states, *i.e.* that the *context* of n is satisfied. The context of all actions is thus *state-based*. If this necessary condition is satisfied, then the action is said to be *feasible*. We note that in a partially observed environment, the agents might not know for sure whether the context of any particular action is satisfied. If the agent chooses an infeasible action, then we assume the agent accrues the cost of the action (which may be higher if the action is chosen when the context is not satisfied than if the context were satisfied), and the action does nothing to affect the state of the graphs.

For example, assume a node n in the precedence graph is in state 0 if it is a non-game node or state $\mathcal{D}$ if it is a game node and the attacker would like to change the state of this node to $\mathcal{A}$. A necessary condition for success (or for an action to be feasible) might be that at least one of the immediate ancestor nodes of n be in state $\mathcal{A}$. Another example of feasibility is that there is a path of ancestor nodes from the start node to node n where every node on this path is currently in state $\mathcal{A}$. This latter example would model the need for a supply chain to be under complete control of the attacker before any further attacks are allowed (or are prudent). We assume the attacker cannot successfully attack a game node if the game node is in state 0.

Thus, both agents take control of non-game nodes in their respective graphs in an identical fashion and without the interference of the other agent (an assumption that is easily modified). However, taking control of game nodes is different for the two agents. The defender must take control of a game node before the node can exist. Once a game node exists, it is initially controlled by the defender, but then becomes a candidate for control by either agent. Hence, the existing game nodes represent potential battlegrounds for the two agents. We depict this existence phenomena in Figure 4.1.

Assume that existing game node n is under the control of one of the two agents. At each epoch, the agent controlling node n can decide to protect node n from an attempt by the other agent to assume its control or to leave node n unprotected. If node n is left unprotected and the other agent attempts to take control, then there is a probability that the other agent will be successful, assuming the attempt is feasible. If node n is protected and the other agent attempts to take control of node n, we assume that the likelihood of success by the other agent is reduced.

Finally, we depict how different types of actions may be employed by the defender to impede attacker progress in Figure 4.2, and remark that this subsection describes one of many possible transition structures for the state of the precedence graph. In Appendix C.1, we show how the transition probabilities $P[S_{t+1}|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}]$ may be computed under this transition structure, including the three types of actions presented above. In the remainder of the paper, however, we assume for the sake of simplicity of exposition, that Take moves are the only type of actions available.

Information Patterns. An information pattern describes what an agent knows before the agent selects an action, *i.e.*, *who knows what and when*. We assume the information pattern for both agents is node specific and can involve data about the state of any node in either graph. We also assume that an agent can not only take action to obtain control of, or continue to control, a node (at a cost) but can also decide to acquire data (at a cost) about the state of individual nodes and that these data may be inaccurate (noise corrupted) and

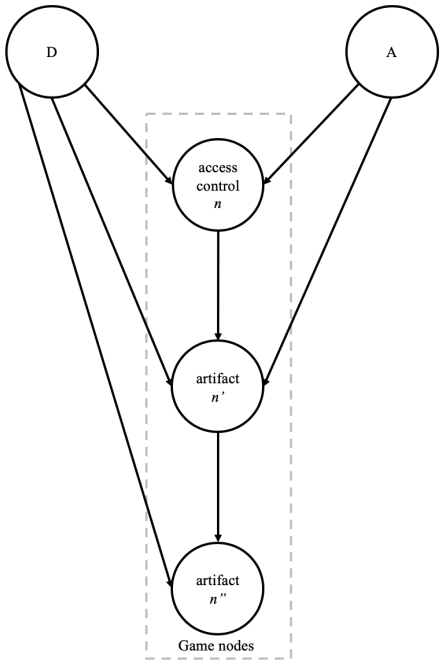


Figure 4.2: *Types of defense moves.* Suppose the attacker has control of n and n'. The defender may seek to prevent the attacker from progressing in the precedence graph in three different ways.

- *Take on n from D.* This corresponds to an attack on the context of the attacker. If the access control is necessary for attacker take moves on n' and n'', then the defender can make these moves infeasible by controlling n.
- *Take on n' from D.* This is a direct attempt to wrest control of n' from the attacker.
- *Take on n'' from D and protect.* The defender takes the subsequent node n'' and employs "protect" actions on n'' to limit the chance of the attacker ever taking control of n''.

can be in either graph. It seems intuitive that more accurate observation quality will result in better expected performance, and this intuition, although not true in general ([62], [9]), will be supported by examples presented later in the paper. Let $\mathscr{I}_t^{\mathcal{A}}$ and $\mathscr{I}_t^{\mathcal{D}}$ be the data available to the attacker and defender, respectively, at epoch t on which to base their action selections. Examples include:

- $\mathscr{I}_t^{\mathcal{A}} = S_t^{\mathcal{A}}$; *i.e.*, the attacker has complete visibility of the state of the precedence graph without cost but only finite memory

- $\mathscr{I}_t^{\mathcal{D}} = Z_t^{\mathcal{D}}$; *i.e.*, the defender has inaccurate observations of the state of the development process graph without cost and only finite memory, where $Z_t^{\mathcal{D}} = \{z_t(n) : n \in N_{\mathcal{D}}\}$ and $z_t(n)$ is a noise corrupted observation of the state of node n just before epoch t.

Detection. We remark that $\mathscr{I}_t^{\mathcal{D}}$ can be useful for attack detection. The probability $P[S_t^{\mathcal{D}} \neq S_0^{\mathcal{D}} | \mathscr{I}_t^{\mathcal{D}}]$ provides insight into whether or not an attack is underway, and the probabilities $P[S_t^{\mathcal{D}} | \mathscr{I}_t^{\mathcal{D}}]$ for all $S_t^{\mathcal{D}}$ can be useful in determining a best response action.

Reward, Cost, Criterion Structure, Problem Objective, and Policies. We assume that any action A_t^k taken by an agent k at epoch t accrues a cost, $c^k(A_t^k)$. We assume that any reward accrued is dependent on the current state of the agent's graph, and denote this reward by $r^k(S_t^k)$. In other words, we assume that the cost function $C^k(S_t, A_t) = c^k(A_t^k) - r^k(S_t^k)$. Thus, the dependence of the cost function for each agent is separable and is only dependent on the actions of the other agent through the state of the agent's respective graph, S_t^k.

As in the POMG, we assume that the criterion each agent is trying to minimize is the expected value of the sum over all epochs of all costs minus the rewards accrued at each epoch, as in Equation 4.1. Recall that each agent is trying to minimize its criterion over all feasible policies, where a policy is a mapping at each epoch from the data set available to the agent at that epoch into the set of actions available to the agent at that epoch. A policy that achieves the minimized criterion is an *optimal* policy. Since each agent's criterion is dependent on the policy taken by the other agent, an agent's optimal policy depends on what policy is taken by the other agent.

We have several remarks at this point:

- A policy can be viewed as a set of IF-THEN rules, stating that if agent k knows $\mathscr{I}_t^k$ at epoch t and has feasible action set $\mathscr{A}_t^k$ at epoch t, then the policy tells the agent what action(s) in $\mathscr{A}_t^k$ to select. A policy answers the question: Given I know $\mathscr{I}_t^k$ and I can select an action from $\mathscr{A}_t^k$, which action should I select? For example, given I'm in my car at the intersection of Peachtree Road and North Avenue headed North,

there is congestion on I-75, and my destination is Lenox Mall, should I go left, right, or straight?

- A policy can contain stochastic rules; *e.g*, if $\mathscr{I}_t^k$ is the current data set for agent k, and if $a, a' \in \mathscr{A}_t^k$, then a stochastic IF-THEN rule might be: select action a with probability p; otherwise, select action a'.

- If an agent has multiple objectives, then instead of seeking an optimal policy, the agent may seek a Pareto optimal frontier with its associated Pareto optimal policies.

4.4 Preliminary Results

Now, we formally specify the POMG model that we consider for the remainder of this paper. Recall that the motivating PRESTIGE problem setting is inherently defender-centric. Namely, the risk analysis that we perform is from the defender's point of view. It is natural, then, that in this practical setting we understand the information pattern more clearly for the defender than the attacker. For instance, it may be difficult to ascertain the quality of the observations received by the attacker, even though we are able to specify this for the defender. Moreover, it is intuitively reasonable (but perhaps not always the case – a topic for future investigation) that if a defender policy is 'good' against an attacker with complete observability of the system, then it will be 'good' against an attacker with partial observability of the system. Since this is the problem setting, we consider there to be an asymmetric information structure between the defender $\mathcal{D}$ and the attacker $\mathcal{A}$. Namely, we assume:

$$\mathscr{I}_t^{\mathcal{D}} = \left\{ Z_t, \ldots, Z_1, A_{t-1}^{\mathcal{D}}, \ldots, A_0^{\mathcal{D}}, \pi^{\mathcal{A}} \right\}$$
$$\mathscr{I}_t^{\mathcal{A}} = \left\{ S_t, \ldots, S_0, A_{t-1}^{\mathcal{A}}, \ldots, A_0^{\mathcal{A}}, \pi^{\mathcal{D}} \right\}.$$

We assume that each agent knows the structure of the respective graphs $(N_{\mathcal{D}}, E_{\mathcal{D}})$ and $(N_{\mathcal{A}}, E_{\mathcal{A}})$.

To simplify the analysis, we assume that the attacker *presumes* that the defender has complete observability of the system, *i.e.* has a symmetric information pattern. In this case, results in [37] guarantee that it is sufficient for the attacker to consider only Markovian deterministic policies, such that $\mathscr{I}_t^A = \{S_t, \pi^{\mathcal{D}}\}$ and $\pi^{\mathcal{A}}$ is a mapping from all possible graph states S_t into feasible actions $\mathscr{A}_t^{\mathcal{A}}$ for all t. The attacker assumes that the defender's policy is likewise Markovian, and thus the attacker can determine the best response policy by solving a Markov decision process (MDP):

$$V_t^{\mathcal{A}}(S_t|\pi^{\mathcal{D}}) = \min_{A_t^{\mathcal{A}} \in \mathscr{A}_t^{\mathcal{A}}} \left\{ C^{\mathcal{A}}\big(S_t, A_t^{\mathcal{A}}, \pi^{\mathcal{D}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1}|S_t, A_t^{\mathcal{A}}, \pi^{\mathcal{D}}(S_t)\big] V_{t+1}^{\mathcal{A}}(S_{t+1}|\pi^{\mathcal{D}}) \right\}. \tag{4.2}$$

Well-known results in [37] guarantee that we may solve this MDP by backward recursion.

Due to the information asymmetry, computing the defender's best response policy is more complicated. Given the attacker policy, the defender may theoretically determine the best response policy by solving the following stochastic dynamic program (DP):

$$V_t^{\mathcal{D}}(\mathscr{I}_{t+1}|\pi^{\mathcal{A}}) = \min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{A}}} \mathbb{E}\left[C_t^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big) + V_{t+1}^{\mathcal{D}}(\mathscr{I}_{t+1}^{\mathcal{D}}|\pi^{\mathcal{A}})|\mathscr{I}_t^{\mathcal{D}}, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}\right]. \tag{4.3}$$

The state space of this formulation grows with t, which makes this formulation unrealistic for computation. Note that this best response DP is a partially observable Markov decision process (POMDP). Results in [45] guarantee that the Bayesian belief distribution over the state space (the union of the development process and precedence graphs) $X_t^{\mathcal{D}} = \{P[S_t|\mathscr{I}_t^{\mathcal{D}}]\}$ is a sufficient statistic for control and that the defender may determine

the best response policy by solving:

$$V_t^{\mathcal{D}}(X_t^{\mathcal{D}}|\pi^{\mathcal{A}}) = \min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{A}}} \Bigg\{ \sum_{S_t} X_t^{\mathcal{D}}(S_t) C_t^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big) + \sum_{S_{t+1}, Z_{t+1}} \sigma\big(Z_{t+1}|X_t^{\mathcal{D}}, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big) V_{t+1}^{\mathcal{D}}\left(\Lambda\big(Z_{t+1}, A_t^{\mathcal{D}}, X_t^{\mathcal{D}}\big)|\pi^{\mathcal{A}}\right) \Bigg\}, \tag{4.4}$$

where

$$\sigma\big(Z_{t+1}|X_t^{\mathcal{D}}, A_t^{\mathcal{D}}, \pi_t^{\mathcal{A}}(S_t)\big) = P\big[Z_t|\mathscr{I}_t^{\mathcal{D}}, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big]$$
$$\Lambda(S_{t+1}|Z_{t+1}, A_t^{\mathcal{D}}, X_t^{\mathcal{D}}) = P\big[S_{t+1}|Z_{t+1}, A_t^{\mathcal{D}}, X_t^{\mathcal{D}}\big].$$

Thus, $X_{t+1}^{\mathcal{D}} = \Lambda(Z_{t+1}, A_t^{\mathcal{D}}, X_t^{\mathcal{D}})$ is the defender's Bayesian posterior belief distribution over the state space, given that the observation was Z_{t+1}, the defender chose action $A_t^{\mathcal{D}}$, and the prior belief was $X_t^{\mathcal{D}}$. We show how to compute Λ using Bayes' rule in the next section.

4.5 Heuristic Solution Procedure

A candidate solution to the POMG of Section 4.4 is a tuple of policies $(\pi^{\mathcal{A}}, \pi^{\mathcal{D}})$. In a standard POMG, ideally these policies would form a Nash equilibrium, in which neither agent has incentive to deviate from their policy. Due to the size and complexity of our problem, determining a Nash equilibrium is computationally and analytically intractable for all but the smallest of problem instances. Moreover, in our application setting the defender policy generated by such a Nash equilibrium may not even be the most desirable to *implement* in practice. Recall that in our application, we assume that the modeler (the defender) is trying to analyze risk in the system against *potential* or *hypothetical* attackers. The objectives of the attacker are assumed or posited by the defender in order to capture different potential attacker priorities (*e.g.* budget-constrained, win-at-all-costs). A Nash equilibrium solution would assume that the objectives of such a potential attacker are well-known and, thus, the attacker has no reason to deviate from the equilibrium if both agents

are behaving completely rationally (an assumption that itself might not hold). Moreover, it is unknown whether the equilibrium is stable or not, so it is unclear whether the defender's equilibrium policy is effective against a *range* of attacker policies.

For these reasons, we use the POMG as a *generative mechanism*. The modeler/defender might posit different objective functions for the attacker in order to generate different kinds of attacker behavior. Moreover, as we showed in Section 4.4, it provides a mechanism for determining best response policies using dynamic programming methods. The focus for the defender is then to determine a "good" policy, *i.e.* a policy that is:

- *robust* to a range of attacker policies,
- *adaptive* to the changing state of the system, and
- *intelligent* in the way that it processes (possibly noise-corrupted) information.

We foreshadowed our approach to policy determination in Section 4.4, where we showed how determining the attacker's best response policy, given a Markovian defender policy, reduces to solving a MDP. Likewise, determining the defender's best response policy, given an attacker policy, reduces to solving a POMDP. Although solving the POMDP defined by Equation 4.4 is a reduction in computational complexity relative to the POMG, it may still be computationally difficult to solve itself. This motivates the following three-fold method for generating a good (randomized) heuristic defender policy:

1. *Train the Agents.* First, we relax the partial observability assumption of the defender and generate intelligent, plausible attacker and defender policies by solving successive best response MDPs.
2. *Generate a Robust Defender Policy.* After training, and again relaxing the partial observability assumption, we solve for a single defender policy by solving a robust MDP that considers all of the trained attacker policies.

3. *Probability Matching Heuristic.* The defender then uses a probability matching heuristic policy based on Thompson sampling.

We detail our heuristic solution procedure in Figure 4.3 and discuss each step in the remainder of this section.

4.5.1 Training the Agents

The first step in our heuristic solution procedure is to "train" the agents. This training step is our mechanism for generating hypothetical attacker policies that exemplify different attacker behaviors. The goal is to generate plausible, intelligent, and adaptive attacker policies that the defender should consider in constructing its own policy. In this step, we relax the partial observability condition for the attacker both because of potential computational intractability — recall the defender's best response problem is a POMDP, which may be difficult to solve — and because it matches the attacker's *presumption* about the defender's information pattern. Thus, given an attacker policy $\pi^{\mathcal{A}}$, we assume that the defender may generate a Markovian deterministic best response policy, $\pi^{\mathcal{D}}$, (a function from S_t to $\mathscr{A}_t^{\mathcal{D}}$ for all t) by solving the following MDP:

$$V_t^{\mathcal{D}}(S_t) = \min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{D}}} \left\{ C^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1} | S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big] V_{t+1}^{\mathcal{D}}(S_{t+1}) \right\}.$$

We say a policy π^k for agent k is *trained* against a policy π^j for agent j, where $j \neq k$ and $k, j \in \{\mathcal{A}, \mathcal{D}\}$, if (and only if) π^k is a best response policy to π^j.

The training procedure begins with an initial finite set of defender policies, $\mathcal{G}_0^{\mathcal{D}}$, which we call "generation-0" defender policies. These policies are initialized by the defender and can contain a wide range of easily determined policies, *e.g.* a naive policy such as "take action A at every epoch and every state", a myopic policy that is greedy with respect to the single-period cost function $C^{\mathcal{D}}$, or the historical operating policies.

Next, we determine the generation-1 attacker policies, $\mathcal{G}_1^{\mathcal{A}}$, by training the attacker

0. *Initialization.* Defender has an initial finite set of defender policies $\mathcal{G}_0^{\mathcal{D}}$ (*i.e.* the generation-0 defender policies).

1. *Train the Agents.* For each generation $g = 1, \ldots, m$:

   (a) For each $\pi^{\mathcal{D}} \in \mathcal{G}_{g-1}^{\mathcal{D}}$ (or any subset of $\mathcal{G}_{g-1}^{\mathcal{D}}$), determine the attacker best response by solving via backward recursion as in [37] for $t = T+1, \ldots, 0$:

$$V_t^{\mathcal{A}}(S_t) = \min_{A_t^{\mathcal{A}} \in \mathscr{A}_t^{\mathcal{A}}} \left\{ C^{\mathcal{A}}\big(S_t, A_t^{\mathcal{A}}, \pi^{\mathcal{D}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1}|S_t, A_t^{\mathcal{A}}, \pi^{\mathcal{D}}(S_t)\big] V_{t+1}^{\mathcal{A}}(S_{t+1}) \right\}.$$

   Add the best response policy, $\pi^{\mathcal{A}}$, to the set of generation-g attacker policies, $\mathcal{G}_g^{\mathcal{A}}$:

$$\mathcal{G}_g^{\mathcal{A}} \leftarrow \mathcal{G}_g^{\mathcal{A}} \cup \pi^{\mathcal{A}}.$$

   (b) For each $\pi^{\mathcal{A}} \in \mathcal{G}_g^{\mathcal{A}}$ (or any subset of $\mathcal{G}_g^{\mathcal{A}}$), determine the defender best response by solving via backward recursion as in [37] for $t = T+1, \ldots, 0$:

$$V_t^{\mathcal{D}}(S_t) = \min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{D}}} \left\{ C^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1}|S_t, A_t^{\mathcal{D}}, \pi^{\mathcal{A}}(S_t)\big] V_{t+1}^{\mathcal{D}}(S_{t+1}) \right\}.$$

   Add the best response policy, $\pi^{\mathcal{D}}$ to the set of generation-g defender policies, $\mathcal{G}_g^{\mathcal{D}}$:

$$\mathcal{G}_g^{\mathcal{D}} \leftarrow \mathcal{G}_g^{\mathcal{D}} \cup \pi^{\mathcal{D}}.$$

2. *Generate a Robust Defender Policy.* Solve the MMDP with weight distribution $\lambda = [\lambda_1, \ldots, \lambda_n]$ over the set of attacker policies $\{\pi_1^{\mathcal{A}}, \ldots, \pi_n^{\mathcal{A}}\}$ determined by training. Generate the robust policy π_{rob}^{D} using WSU, adapted from [51]. Initialize $\hat{V}_{T+1}^{i} = C_{T+1}^{\mathcal{D}}$ for each policy π_i^{A}, $i = 1, \ldots, n$.

   For $t = T, \ldots, 0$:

   (a) For every state S_t:

$$\pi_{rob}^{\mathcal{D}}(S_t) \leftarrow \arg\min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{D}}} \sum_{i=1}^{m} \lambda_i \left\{ C^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi_i^{\mathcal{A}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1}|S_t, A_t^{\mathcal{D}}, \pi_i^{\mathcal{A}}(S_t)\big] \hat{V}_{t+1}^{i}(S_{t+1}) \right\}.$$

   (b) For every state S_t and $i = 1, \ldots, n$:

$$\hat{V}_t^{i}(S_t) \leftarrow \min_{A_t^{\mathcal{D}} \in \mathscr{A}_t^{\mathcal{D}}} \left\{ C^{\mathcal{D}}\big(S_t, A_t^{\mathcal{D}}, \pi_i^{\mathcal{A}}(S_t)\big) + \sum_{S_{t+1}} P\big[S_{t+1}|S_t, A_t^{\mathcal{D}}, \pi_i^{\mathcal{A}}(S_t)\big] \hat{V}_{t+1}^{i}(S_{t+1}) \right\}.$$

3. *Probability Matching Heuristic.* The defender plays the randomized policy such that at epoch t, $\pi_{rob}^{\mathcal{D}}(S_t)$ is chosen with probability $X_t^{\mathcal{D}}(S_t)$.

Figure 4.3: Heuristic solution procedure.

against the generation-0 defender policies, $\mathcal{G}_0^{\mathcal{D}}$. That is, we begin with $\mathcal{G}_1^{\mathcal{A}} = \varnothing$. Then for each policy $\pi^{\mathcal{D}} \in \mathcal{G}_0^{\mathcal{D}}$, we determine a best response policy $\pi^{\mathcal{A}}$ by solving the attacker MDP induced by $\pi^{\mathcal{D}}$ (Equation 4.2), and add it to the set $\mathcal{G}_1^{\mathcal{A}}$. The set of generation-1 defender policies, $\mathcal{G}_1^{\mathcal{D}}$, are likewise determined by training against the generation-1 attacker policies, $\mathcal{G}_1^{\mathcal{A}}$.

For generation-g policies, $\mathcal{G}_g^{\mathcal{A}}$ is the set of attacker policies trained against $\mathcal{G}_{g-1}^{\mathcal{D}}$, and $\mathcal{G}_g^{\mathcal{D}}$ is the set of defender policies trained against $\mathcal{G}_g^{\mathcal{A}}$. In this way, the generation number g may be interpreted as the *order of rationality*, *i.e.* the number of times each agent considers the other agent's best response in determining their policy. This is alternatively referred to in the game theoretic literature as *level-g reasoning* or *depth of reasoning* ([50]) and is closely related to the concept of cognitive hierarchy in games ([13]). Thus, given the initial defender policies $\mathcal{G}_0^{\mathcal{D}}$, we generate the sets of level-g reasoned policies for the attacker and the defender, up to some pre-specified maximal level of reasoning, m.

There are other implementations that we might consider that give the generation number g alternative interpretations. For instance, we might implement the training procedure as a genetic algorithm that mimics the natural selection and reproduction process in biological evolution. In this case, we assume we have access to some fitness function, F, that gives each policy a real-numbered "fitness score" (a natural fitness score might be the evaluation of the policy given the opponent policy at the initial state of the system, *e.g.* $V_0^{\pi^{\mathcal{D}}}(S_0|\pi^{\mathcal{A}})$). At generation g, we assume that we have the set of "surviving" defender policies, $\mathcal{G}_{g-1}^{\mathcal{D}}$, and attacker policies, $\mathcal{G}_{g-1}^{\mathcal{A}}$ the fitness score for each of the generation-$(g-1)$ policies satisfied some "survivability" criterion. We might then consider $\mathcal{G}_g^{\mathcal{A}}$ to be the set of best response policies to the surviving defender policies $\mathcal{G}_{g-1}^{\mathcal{D}}$ that satisfy the survivability criterion, and likewise, $\mathcal{G}_g^{\mathcal{D}}$ to be the defender best response policies to $\mathcal{G}_g^{\mathcal{A}}$. In this context, the process of solving the best response MDP for each agent acts as the "genetic operator" for generating the next generation of candidate policies. This process continues for a pre-specified maximal number of generations, m.

4.5.2 Generating a Robust Defender Policy

At the end of the training step, we have a finite number of attacker policies, $\{\pi_1^{\mathcal{A}}, \ldots, \pi_n^{\mathcal{A}}\}$. In this step, we want to generate a single defender policy that is robust against this set of attacker policies, and thus robust against an array of attacker types. In order to do so, we use the multi-model MDP (MMDP) of [51], that was originally developed for imprecisely specified MDPs in which there are multiple candidate models of the cost function and transition probabilities. In our setting, we recognize that each attacker policy $\pi_i^{\mathcal{A}}$ induces a different transition probability structure for the defender $\{P[S_{t+1}|S_t, A_t^{\mathcal{D}}, \pi_i^{\mathcal{A}}(S_t)]\}$, and thus represents a candidate model of the transition probabilities. In the MMDP, we suppose that we have access to a set of weights $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ such that λ_i represents the *a priori* probability that the defender believes the attacker will use policy $\pi_i^{\mathcal{A}}$. A robust defender policy, $\pi_{rob}^{\mathcal{D}}$, is then a policy that minimizes the expected costs, given λ:

$$\pi_{rob}^{\mathcal{D}} \in \arg\min_{\pi^{\mathcal{D}}} \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \sum_{t=1}^{T} C^{\mathcal{D}}\big(S_t, \pi^{\mathcal{D}}(S_t), \pi_i^{\mathcal{A}}(S_t)\big) + C_{T+1}^{\mathcal{D}}(S_{T+1})\right].$$

Standard solution methods in [37] like backward recursion are not applicable for this MMDP. [51] suggest an algorithm called "Weight-Select-Update" (WSU) that we adapt to our setting here in Figure 4.3.

4.5.3 Probability Matching Heuristic

In training and determining $\pi_{rob}^{\mathcal{D}}$ we relaxed the partial observability condition on the defender, so that $\pi_{rob}^{\mathcal{D}}$ is a function that maps states into feasible defender actions. In implementation, or policy evaluation, the defender must make decisions at epoch t on the basis of the information available:

$$\mathscr{I}_t^{\mathcal{D}} = \left\{Z_t, \ldots, Z_1, A_{t-1}^{\mathcal{D}}, \ldots, A_0^{\mathcal{D}}, \pi^{\mathcal{A}}\right\},$$

rather than knowledge of the true state S_t. Recall that $X_t^{\mathcal{D}} = \{P[S_t|\mathscr{I}_t^{\mathcal{D}}]\}$ is a sufficient statistic for the defender. In implementation, we assume the defender's action at epoch t is selected randomly, based only on knowledge of $\mathscr{I}_t^D$, such that at epoch t, the defender chooses the robust policy action for state S_t with the probability that the defender believes the system to be in state S_t. More precisely, in order to differentiate between the true state of the system S_t and the random variable upon which the defender makes a decision, suppose that R_t is a random variable with the same support as S_t, probability mass function $X_t^{\mathcal{D}}$, and is independent of S_t. The defender chooses its action as follows: R_t is drawn from $X_t^{\mathcal{D}}$, then the defender plays $\pi^{\mathcal{D}}(R_t)$. Thus, the defender chooses action $\pi_{rob}^{\mathcal{D}}(S_t)$ with probability $X_t^{\mathcal{D}}(S_t)$.

Note that this is a *probability matching* heuristic. Denote by $\pi_{heu}^{\mathcal{D}}$ this policy, such that $\pi_{heu}^{\mathcal{D}}(X_t^{\mathcal{D}}) = \pi_{rob}^{\mathcal{D}}(S_t)$ with probability $X_t^{\mathcal{D}}(S_t)$. This heuristic policy, $\pi_{heu}^{\mathcal{D}}$, is conceptually connected to Thompson sampling — a popular technique in machine learning for balancing the well-known exploration-exploitation tradeoff in online optimization problems, which we detail in Appendix C.3.

4.5.4 Step-by-step explanation of an example transition from t to $t + 1$.

In this subsection, we consider a hypothetical unified graph in order to elucidate the information flow and transition dynamics from t to $t + 1$. In this unified graph, we assume that each node is initially controlled by the defender (that they already exist) and each node is a game node. For simplicity, we assume that the only actions available to the attacker and defender are Take moves, which we refer to as "attacks" for the attacker and "counter-attacks" for the defender.

We assume that the attacker's policy, $\pi^{\mathcal{A}}$, has been determined and that the defender knows the attacker's policy and has constructed a policy, $\pi^{\mathcal{D}}$, to counteract it in order to best achieve the defender's objectives (protect node 10). The attacker's policy (constructed to compromise node 10) directs the attack from node 1 to node 6 through nodes 4 and 5.

1. *Information patterns*: who knows what at epoch t.
   (a) The attacker knows the state of the precedence and development process graph, *i.e.* $\mathscr{I}_t^{\mathcal{A}} = S_t$.
   (b) The defender keeps track of the belief distribution $X_t^{\mathcal{D}} = P[S_t|\mathscr{I}_t^{\mathcal{D}}]$.
2. *The attacker's action*: given the attacker's policy $\pi^{\mathcal{A}}$, the attacker's action is $\pi^{\mathcal{A}}(S_t)$.
3. *The defender's action*:
   (a) Based on $X_t^{\mathcal{D}}$, a random process generates the realization R_t, which is revealed to the defender.
   (b) Given the defender's policy $\pi^{\mathcal{D}}$, the defender's action is $\pi^{\mathcal{D}}(R_t)$.
4. *Update the precedence graph*: Given $\pi^{\mathcal{A}}(S_t)$ and $\pi^{\mathcal{D}}(R_t)$, the state of the precedence graph is updated to S_{t+1} according to the transition probability $P[S_{t+1}|S_t, \pi^{\mathcal{A}}(S_t), \pi^{\mathcal{D}}(R_t)]$.
5. *Update the belief distribution.* The defender receives observation Z_{t+1} and has information pattern $\mathscr{I}_{t+1}^{\mathcal{D}} = \{Z_{t+1}, R_t, \mathscr{I}_t^{\mathcal{D}}\}$. The defender then updates the belief distribution:
$$X_{t+1}^{\mathcal{D}} \leftarrow \Lambda\left(Z_{t+1}, \pi^{\mathcal{D}}(R_t), X_t^{\mathcal{D}}\right) = \{P\left[S_{t+1}|\mathscr{I}_{t+1}^{\mathcal{D}}\right]\}.$$
6. Update $t \leftarrow t + 1$; GO TO 1.

Figure 4.4: How the process proceeds, from one decision epoch to the next, given attacker policy $\pi^{\mathcal{A}}$ and the defender policy $\pi^{\mathcal{D}}$.

Since the defender knows the attacker's policy, the defender is aware that this will be the line of attack.

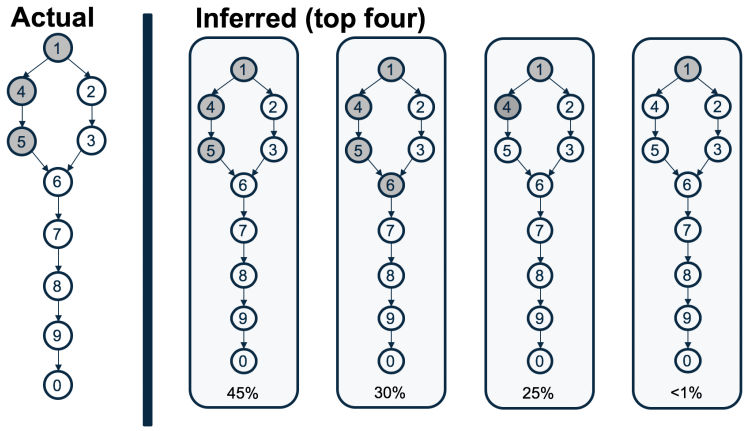


Figure 4.5: Configuration at epoch t.

Figure 4.5. Figure 4.5 summarizes what the agents know at decision epoch t. The actual state of the precedence graph is that nodes 4 and 5 have been compromised $\big($i.e., the state of the precedence graph is $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})\big)$, which the attacker knows. Based on possibly inaccurate observations of the state of the nodes, the defender does not know whether only node 4 has been compromised $\big($state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})\big)$, both nodes 4 and 5 have been compromised $\big($state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})\big)$, or nodes 4, 5, and 6 have been compromised $\big($state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})\big)$. However, the defender has inferred from past and current observations and the attacker's and defender's policies that the probabilities of these three possible states are 0.25, 0.45, and 0.30, respectively.

Figure 4.6. We assume that the defender's policy selects the defender's action on the basis of the current state of the precedence graph. We determine what state to use randomly, selecting state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.25, state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.45, and state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.30. Figure 4.6 indicates that the realization of the random variable making this selection, R_t, selected state $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$, which coincidently is the ac-

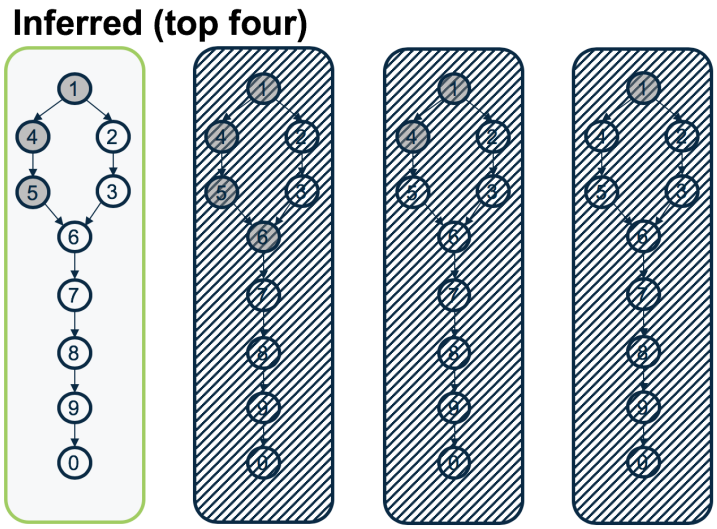


Figure 4.6: A "guess" state, R_t, is randomly chosen by the defender (in green) according to the inference distribution, $\{P[S_t|\mathscr{I}_t^{\mathcal{D}}]\}$.

tual current state of the precedence graph.

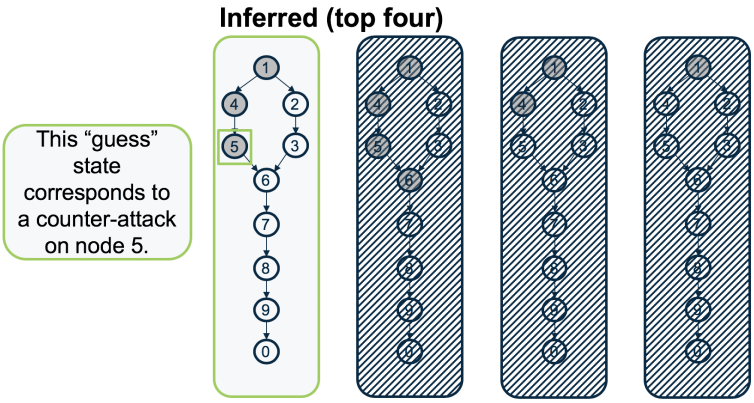


Figure 4.7: This "guess" state, R_t, determines the defender's action to be taken, $\pi^{\mathcal{D}}(R_t)$.

Figure 4.7. Given state $R_t = (\mathcal{A}, \mathcal{D}, \mathcal{D}, \mathcal{A}, \mathcal{A}, \mathcal{D}, \mathcal{D}, \mathcal{D}, \mathcal{D}, \mathcal{D})$, the defender's policy selects the action: counter-attack node 5, which corresponds to $\pi^{\mathcal{D}}(R_t)$.

Figure 4.8. Given the actual state of the precedence graph, S_t, the attacker's policy selects the action: attack node 6 from (compromised) node 5, which corresponds to $\pi^{\mathcal{A}}(S_t)$.

Figure 4.9. Figure 4.9 indicates that the counter-attack was successful. Since we have assumed that the outcome of the defender's action take precedence over the outcome of

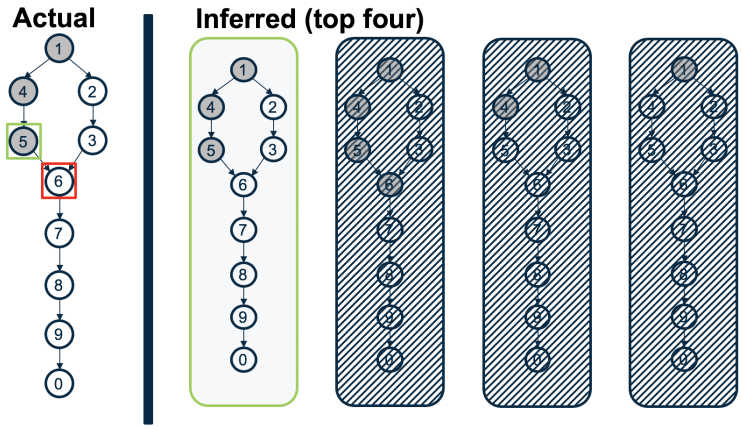


Figure 4.8: On the basis of the attacker policy, $\pi^{\mathcal{A}}$, and the actual state, S_t, the attacker chooses action $\pi^{\mathcal{A}}(S_t)$. On the basis of the defender policy, $\pi^{\mathcal{D}}$, and the inference distribution, $\left\{P\left[S_t|\mathscr{I}_t^{\mathcal{D}}\right]\right\}$, the defender chooses randomized action $\pi^{\mathcal{D}}(R_t)$, where $R_t \sim \left\{P\left[S_t|\mathscr{I}_t^{\mathcal{D}}\right]\right\}$.

the attacker's action, the attacker's action fails. Thus, the state of the precedence graph at epoch $t+1$ becomes $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$, indicating that the attacker has lost ground.

Figure 4.10. The defender now receives new observations of the state of the nodes in the precedence graph. We note two observations are incorrect. The defender is informed that nodes 5 and 6 are compromised when they both are not. These incorrect data are used to update the probabilities on the states of the precedence graph for the defender.

Figure 4.11. Figure 4.11 summarizes what the agents know at decision epoch $t+1$. The actual state is $S_{t+1} = (\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$, which is known by the attacker, and the defender has inferred that the state is $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.20, $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.30, $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.40, and $(\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{A},\mathcal{A},\mathcal{A},\mathcal{A},\mathcal{D},\mathcal{D},\mathcal{D})$ with probability 0.10.

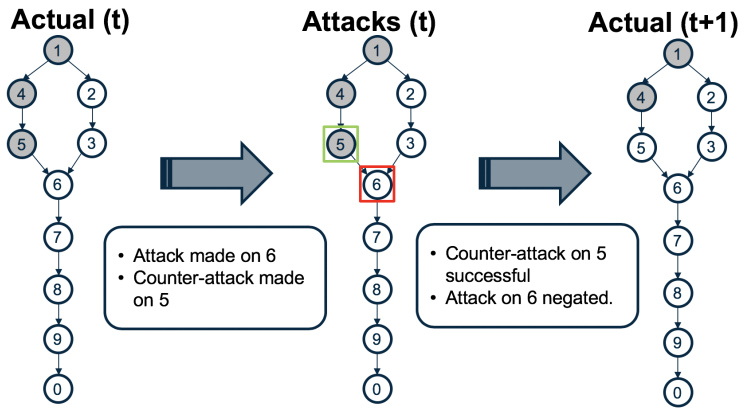


Figure 4.9: State transition occurs on the basis of the actions taken. This transition is fully observed by attacker, but not by defender.

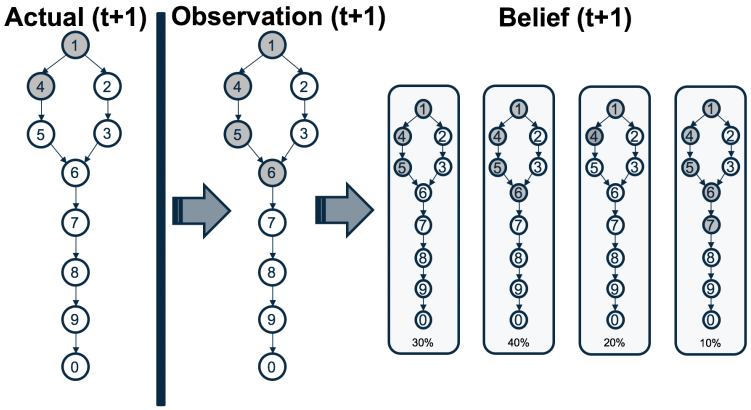


Figure 4.10: Defender gets a noisy observation of the new state, Z_{t+1}, and updates his belief distribution to $\left\{P\left[S_{t+1}|\mathscr{I}^{\mathcal{D}}_{t+1}\right]\right\}$ (recall that $\mathscr{I}^{\mathcal{D}}_{t+1} = \{Z_{t+1}, R_t, \mathscr{I}^{\mathcal{D}}_t\}$).

4.6 Numerical Exemplar

4.6.1 Set Up

We now present a numerical exemplar to demonstrate the types of insights that our POMG model and solution procedure can generate.

State and Actions. The state of our numerical exemplar is defined by the precedence graph in Figure 4.12a. We assume that the defender has knowledge of the entirety of the precedence graph, such that the development process graph consists of a dummy node with a directed edge emanating to each node 2 through 8. Thus, the attacker and defender

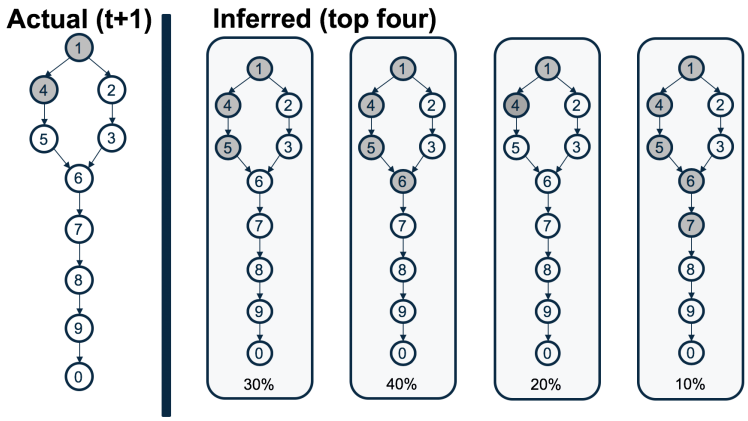
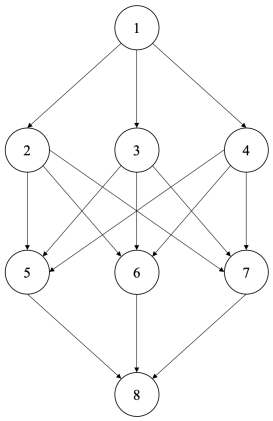


Figure 4.11: Configuration at epoch $t + 1$.



(a) Precedence graph for the numerical exemplar.

Node	$P^{\mathcal{D}}$
2	0.568
3	0.522
4	0.817
5	0.524
6	0.638
7	0.305
8	0.635

(b) The success probabilities for defender take moves.

Node From	Node To	$P^{\mathcal{A}}$
1	2	0.918
1	3	0.490
1	4	0.667
2	5	0.857
2	6	0.189
2	7	0.985
3	5	0.021
3	6	0.169
3	7	0.404
4	5	0.361
4	6	0.886
4	7	0.481
5	8	0.693
6	8	0.353
7	8	0.306

(c) The success probabilities for attacker take moves.

Figure 4.12: Model inputs for the numerical exemplar.

interact on a common set of nodes. For simplicity we assume that the attacker and the defender only have access to "take" moves. Further, the attacker is permitted to perform a take move on one node at each decision epoch, while the defender can perform any number of take moves.

Transition probabilities. The transition probabilities are defined as in Section 4.3, with the probability of success for each node-specific take move defined by Figures 4.12b and 4.12c. In order to generate an additional strategic consideration for the agents, we assume that the probability of success for a take move by the attacker is scaled by the proportion of nodes that the attacker controls in the precedence graph. This makes the probability of

success state-dependent and incentivizes the attacker to consider tradeoffs between short-term progress and the added benefit of long-term increased probability of success. The transition probabilities may be computed as in Appendix C.1, with each term $P^{\mathcal{A}}\big(\tau_t^{\mathcal{A}}(n)\big)$ multiplied by $\frac{1}{|N_{\mathcal{A}}|}\sum_{n\in N_{\mathcal{A}}} \mathbf{1}\{S_t^{\mathcal{A}}(n) = \mathcal{A}\}$.

Observations. $Z_t|S_t$ is assumed to be distributed such that $P[Z_t = s|S_t = s] = p$ and $P[Z_t = s'|S_t = s] = \frac{1-p}{\#\text{ of states}-1}$ for $s \neq s' \in \mathcal{S}$. That is, the parameter p indicates the probability that the defender gets an observation that corresponds to the true state of the system, and an observation corresponding to an incorrect state is uniformly distributed. Note that $p = 1$ corresponds to complete observability and $p = \frac{1}{\#\text{ of states}}$ corresponds to complete unobservability of the system.

Cost functions. The cost function for the defender is specified such that each take move, $\tau^{\mathcal{D}}(n)$ costs $10 \cdot P^{\mathcal{D}}\big(\tau^{\mathcal{D}}(n)\big)$. The terminal cost function for the defender is $C_{T+1}^{\mathcal{D}}(S_t^{\mathcal{D}}) = 1000 \cdot \mathbf{1}\{S_t^{\mathcal{D}}(n = 8) = \mathcal{A}\}$. Further, if the state of node 8 is $\mathcal{A}$, then the defender accrues an additional cost of 1000. We consider four different "types of attackers" — each with different cost functions.

1. *Win only.* This attacker only cares about "winning" at the end of the time horizon, *i.e.* controlling node 8 at epoch $T + 1$. That is, the attacker accrues a cost of -1000 if $S_{T+1}^{\mathcal{A}}(n = 8) = \mathcal{A}$. Otherwise, the attacker accrues no cost.

2. *Win early.* This attacker only cares about controlling node 8 for as long as possible. That is, the attacker accrues a cost of -1000 if $S_t^{\mathcal{A}}(n = 8) = \mathcal{A}$ for any epoch t. Otherwise, the attacker accrues no cost.

3. *Progressive.* This attacker cares about controlling node 8 and other nodes for as long as possible. That is, the attacker accrues a cost of -1000 if $S_t^{\mathcal{A}}(n = 8) = \mathcal{A}$ for any epoch t, and additionally accrues a cost of -100 if $S_t^{\mathcal{A}}(n) = \mathcal{A}$ for any epoch t and node $n \in \{1, 2, \ldots, 7\}$.

4. *Detection skittish.* This attacker is sensitive to "losing ground". That is, the attacker

		$\pi^{\mathcal{D}}_{1,1}$	$\pi^{\mathcal{D}}_{1,2}$	$\pi^{\mathcal{D}}_{1,3}$	$\pi^{\mathcal{D}}_{1,4}$	$\pi^{\mathcal{D}}_{2,1}$	$\pi^{\mathcal{D}}_{2,2}$	$\pi^{\mathcal{D}}_{2,3}$	$\pi^{\mathcal{D}}_{2,4}$
p	0	-35%	-2%	-1%	-36%	-6%	-1%	-2%	-5%
	0.25	-44%	1%	1%	-45%	-6%	1%	2%	-6%
	0.5	-56%	1%	0%	-57%	-14%	-2%	1%	-11%
	0.75	-64%	0%	-3%	-65%	-20%	-1%	-1%	-14%
	1	-70%	2%	2%	-71%	-25%	2%	1%	-18%

Figure 4.13: *Value of Robustness.* For each parameter p, the relative objective value (averaged over all generated attack policies) of $\pi^{\mathcal{D}}_{rob}$ compared to $\pi^{\mathcal{D}}_{i,g}$, where $\pi^{\mathcal{D}}_{i,g}$ indicates the g-generation policy trained against the g-generation attacker policy of type i. For example, the average objective value across all attacker policies of $\pi^{\mathcal{D}}_{rob}$ is 35% better than $\pi^{\mathcal{D}}_{1,1}$.

		$\pi^{\mathcal{A}}_{1,1}$	$\pi^{\mathcal{A}}_{1,2}$	$\pi^{\mathcal{A}}_{1,3}$	$\pi^{\mathcal{A}}_{1,4}$	$\pi^{\mathcal{A}}_{2,1}$	$\pi^{\mathcal{A}}_{2,2}$	$\pi^{\mathcal{A}}_{2,3}$	$\pi^{\mathcal{A}}_{2,4}$
p	0	0%	0%	0%	0%	0%	0%	0%	0%
	0.25	-17%	-27%	-24%	-20%	-50%	-26%	-22%	-43%
	0.5	-37%	-51%	-50%	-43%	-71%	-50%	-49%	-74%
	0.75	-54%	-66%	-66%	-60%	-83%	-65%	-67%	-86%
	1	-60%	-75%	-76%	-67%	-88%	-75%	-78%	-90%

Figure 4.14: *Value of Information.* Relative change in objective value under the robust policy, $\pi^{\mathcal{D}}_{rob}$ across observation parameters, p, against the generated attacker policies.

is penalized with a cost of 100 for each node that switches from $\mathcal{A}$ to $\mathcal{D}$ from epoch t to $t+1$, and accrues a cost of -100 for each node that switches from $\mathcal{D}$ to $\mathcal{A}$. We note that the cost accrued for the detection skittish attacker depends on the state transition, and so we consider the cost function for each decision epoch to be the expected cost over possible state transitions.

Evaluation. We generate policy pairs as in Section 4.5, beginning with the defender policy, $\pi^{\mathcal{D}}_{ACA}$, in which the defender "always counter-attacks", *i.e.* the defender performs a take move on every node $\{2, \ldots, 8\}$ at each epoch. We consider policies of one and two orders of rationality, and depict the policies by $\pi^{\mathcal{D}}_{i,g}$ to represent a defender trained against attacker type i and order of rationality g. The attacker policies are likewise defined as $\pi^{\mathcal{A}}_{i,g}$ for the attacker of type i, trained against $\pi^{\mathcal{D}}_{i,g-1}$. If $g - 1 = 0$, then this corresponds to the attacker of type i trained against $\pi^{\mathcal{D}}_{ACA}$. We evaluate all policy pairs, under the defender heuristic defined in Section 4.5, with 2500 Monte Carlo simulations and a horizon $T = 19$.

	$\pi^{\mathcal{D}}_{1,1}$	$\pi^{\mathcal{D}}_{1,2}$	$\pi^{\mathcal{D}}_{1,3}$	$\pi^{\mathcal{D}}_{1,4}$	$\pi^{\mathcal{D}}_{2,1}$	$\pi^{\mathcal{D}}_{2,2}$	$\pi^{\mathcal{D}}_{2,3}$	$\pi^{\mathcal{D}}_{2,4}$	$\pi^{\mathcal{D}}_{rob}$
$\pi^{\mathcal{A}}_{1,1}$	118	162	163	113	180	147	153	173	170
$\pi^{\mathcal{A}}_{1,2}$	3,133	997	1,018	3,320	1,406	996	995	1,196	984
$\pi^{\mathcal{A}}_{1,3}$	3,449	870	917	3,540	1,242	882	897	1,154	934
$\pi^{\mathcal{A}}_{1,4}$	92	123	128	82	150	132	124	150	131
$\pi^{\mathcal{A}}_{2,1}$	513	191	192	438	140	185	195	202	191
$\pi^{\mathcal{A}}_{2,2}$	3,202	922	978	3,327	1,404	957	972	1,237	1,013
$\pi^{\mathcal{A}}_{2,3}$	3,489	896	789	3,518	1,198	858	892	1,146	836
$\pi^{\mathcal{A}}_{2,4}$	548	117	115	553	130	121	113	74	112

Figure 4.15: Average of Monte Carlo simulated defender objective values for each policy pairing and $p = 1$.

4.6.2 Results

The numerical output of the Monte Carlo evaluations of each of the generated attacker-defender policy pairs is summarized in Figures 4.13, 4.14, and 4.15. Our numerical results highlight two phenomena, in particular — the value of robustness and the value of information.

Recall that our motivation for developing a defender heuristic on the basis of $\pi^{\mathcal{D}}_{rob}$ was based upon the conjecture that a policy trained against a *specific* attacker policy may be subject to vulnerabilities when the attacker chooses a different policy. We use a Monte Carlo simulation of the robust policy against the generated attacker policies (as in Section 4.4) and compare this to the performance of the variously generated defender policies in order to investigate this conjecture. The average costs accrued for each attacker-defender policy pairing are depicted in Figure 4.15 for the completely observed case ($p = 1$). On the whole, we see that the robust policy tends to perform nearly as well as the best defender policy for each attacker policy. In Figure 4.13, we depict the average simulated cost across all attacker policies for each defender policy, relative to the robust policy, at differing levels of observation quality. This shows that mismatched policies can lead to poor performance, *e.g.* for the completely observed case ($p = 1$), $\pi^{\mathcal{D}}_{rob}$ accrued an average of 70% less cost than $\pi^{\mathcal{D}}_{1,1}$ across the different generated policies. We call this phenomenon, in which the robust policy hedges against the risk of policy mismatch, the *value of robustness*.

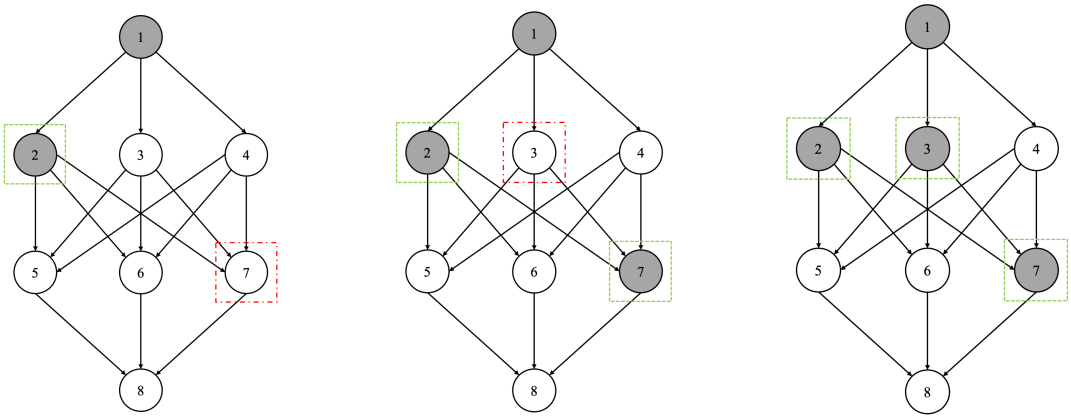


Figure 4.16: *Opener.* Actions under $\pi^{\mathcal{A}}_{1,1}$ are depicted in red, actions under $\pi^{\mathcal{D}}_{1,1}$ are depicted in blue, and actions under $\pi^{\mathcal{D}}_{rob}$ are depicted in green.

It is intuitive that the better the information quality available to the defender ($p \to 1$), the better the defender will perform — a phenomenon we call the *value of information*. (Although this intuition is correct when the policy used is optimal and is often the case for good sub-optimal policies, see [62] for counterexamples.) For the exemplar and policies that we consider here, costs decrease as p gets closer to 1, as summarized in Figure 4.14. Further, we see that better information leverages the value of robustness. As p approaches 1, in Figure 4.13, the robust policy in general performs better relative to the other defender policies. Somewhat counter-intuitively, this may be because the observational uncertainty when p is low can act as a hedge against conservative tendencies for mismatched policies (the defender choosing not to act due to poor assumptions about the attacker, say when $\pi^{\mathcal{D}}_{1,1}$ is trained against a conservative attacker and matched with $\pi^{\mathcal{A}}_{2,3}$, an aggressive attacker).

4.6.3 Closer Examination of Robustness

In this subsection, we compare a specifically trained policy and the robust policy to better understand how the robust policy hedges against the risk of alternative attacker policies in its decision-making. Consider the attacker policy $\pi^{\mathcal{A}}_{1,1}$, the defender policy $\pi^{\mathcal{D}}_{1,1}$, and the robust policy $\pi^{\mathcal{D}}_{rob}$. We depict in Figures 4.16, 4.17, and 4.18 the decisions of these policies under different scenarios that correspond to the attack line — the actions taken by

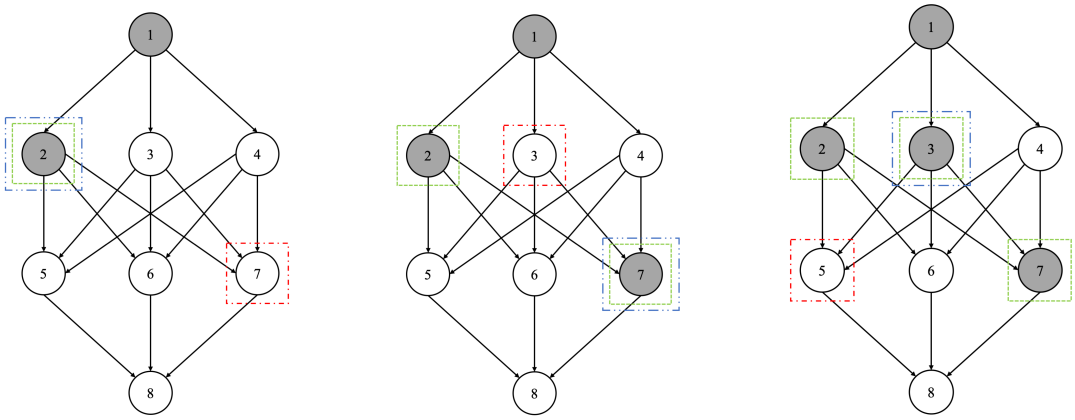


Figure 4.17: Middle game. Actions under $\pi^{\mathcal{A}}_{1,1}$ are depicted in red, actions under $\pi^{\mathcal{D}}_{1,1}$ are depicted in blue, and actions under $\pi^{\mathcal{D}}_{rob}$ are depicted in green.

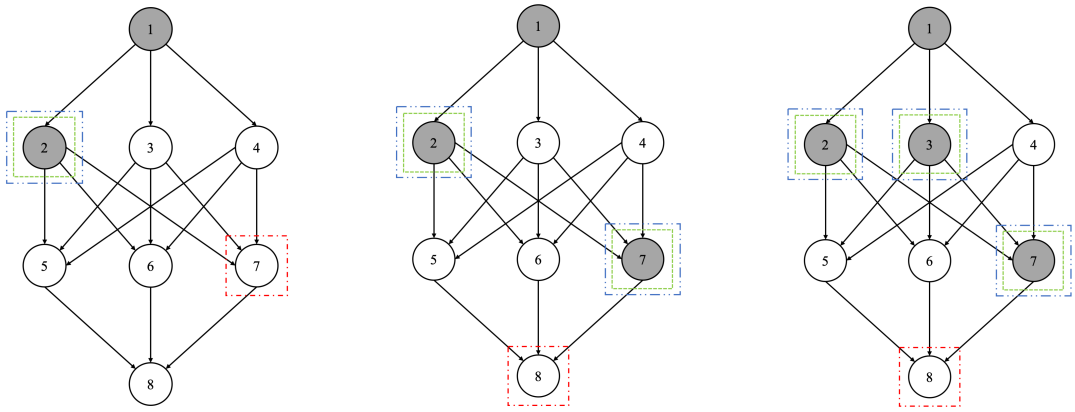


Figure 4.18: End game. Actions under $\pi^{\mathcal{A}}_{1,1}$ are depicted in red, actions under $\pi^{\mathcal{D}}_{1,1}$ are depicted in blue, and actions under $\pi^{\mathcal{D}}_{rob}$ are depicted in green.

the attacker if each attack is successful — under policy $\pi^{\mathcal{A}}_{1,1}$.

Recall that attacker type 1 is "win only", so the attacker playing $\pi^{\mathcal{A}}_{1,1}$ is only concerned with controlling node 8 at the end of the horizon. The attack line for policy $\pi^{\mathcal{A}}_{1,1}$ is $1 \to 2 \to 7 \to 3 \to 8$. The game plays out roughly in three stages — the opener (roughly epochs 1 through 13), the mid game (roughly epochs 13-16), and the end game (roughly epochs 17-19). In the opener and mid game, the attacker seeks to build his or her position, to set his- or herself up for a take from 7 on 8 in the end game. That is, the attacker under $\pi^{\mathcal{A}}_{1,1}$ chooses to build a position and then wait or secure the position until the end of the horizon, at which point the attacker takes on 8. Under the assumption of this behavior, we see that the defender under policy $\pi^{\mathcal{D}}_{1,1}$ chooses to conserve resources in the opener, employ

conservative take moves in the middle game, and play aggressively in the end game. On the other hand, the robust policy plays like the $\pi^{\mathcal{A}}_{1,1}$ in the end game, but for the entirety of the horizon.

Herein lies the value of robustness in this exemplar, the robust policy spreads counter-attacks across multiple nodes in order to cover for different attack lines (spatial robustness), and also attacks early and consistently (temporal robustness). Because the defender under $\pi^{\mathcal{D}}_{1,1}$ does not contest the attacker early in the game, an aggressive attacker can exploit this behavior. The robust policy hedges against this risk by playing aggressively across each decision epoch. This also explains why the policies trained against more aggressive attacker types, $\{\pi^{\mathcal{D}}_{1,2}, \pi^{\mathcal{D}}_{1,3}, \pi^{\mathcal{D}}_{2,2}, \pi^{\mathcal{D}}_{2,3}\}$, perform similarly to the robust policy in this exemplar, as shown in Figure 4.15.

4.7 Conclusions & Future Directions

In this paper, we considered how to create dynamic, data-driven, and robust policies that generate *trust* in development processes. We first presented a graph-based representation of both the development process — necessary conditions for progress towards the defender's end goal — and paths the attacker could possibly take to compromise the defender's end goal (which we call the precedence graph). We then showed in Sections 4.3 and 4.4 that dynamic interactions between the two agents can be modeled using a POMG. In using the POMG as a mechanism for generating varied attacker policies, we were able to construct a defender policy that explicitly considers an array of attacker policies — thereby generating trustworthiness by increasing robustness to misapprehension of attacker objectives and rationality.

We now point out several limitations in our approach that create future research directions. First of all, the graph-based representation of the development process and attacker precedence structure leads to an exponentially increasing state space in the number of nodes comprising the unified graph. This makes our method difficult to implement, as presented

here, for graphs of moderate to large size. In Appendix C.4, we show that the probability transition matrices that govern dynamics are sparse — a fact that we use to significant computational advantage in our exemplar in Section 4.6. Even if exploiting sparsity, we expect computational intractabilities for larger-sized graphs. This opens up substantial opportunity for adapting our robust policy procedure in Figure 4.3 to incorporate approximate dynamic programming and heuristic search methods.

Secondly, in our solution procedure in Section 4.5, we use the POMG as a generative mechanism, using orders of rationality in response to an initial set of defender policies and different posited attacker objectives. Thus, we can trust that the robust defender policy will perform better than a singularly-trained policy against an array of attacker types. However, we recognize that this robustness is contingent upon (1) the degree to which our generative mechanism determines good attacker policies, and (2) the degree to which we have determined plausible attacker objectives. Thus, there is substantial opportunity to improve upon our self-learning approach by improving the generative mechanism. We mention the passing similarity of our method to genetic algorithms (GAs) in Section 5, and note that the evolutionary approach of GAs represents a potentially fruitful research direction.

Finally, we note that we assume that the robust weight parameters, $\lambda = [\lambda_1, \ldots, \lambda_n]$, are chosen *a priori* and are fixed. Instead, we might allow for dynamic tuning of the weight parameters over decision epochs. This would add additional complexity, but will allow for improvement in the robust policy. Further, this would take a step towards making robustness — especially in a reinforcement learning context — *explainable*. For large problems, the analysis we did of robustness in Section 4.6.3 would be difficult due to the size of the problem. If we allow for dynamic adjustment of λ, we can explain robustness by analyzing how the weights considered in the robust policy evolved. Highly weighted attacker policies would indicate significance in forming the robust defender policy.

CHAPTER 5
CONCLUSIONS & FUTURE RESEARCH

There are many potentially fruitful research directions that emanate from each chapter of this dissertation. We detail some of these, below.

5.1 Sequential Decision-Making Affected by Partially Observable Exogenous Forces

In Chapter 2, we consider sequential decision-making environments in which there is a natural hierarchy of effects — micro-level forces that the DM can control, and macro-level forces that the DM cannot. The specially-structured partially observable Markov decision process (POMDP) that we present, the modulated POMDP (M-POMDP), had three main properties that make constructing optimal policies tractable:

1. The M-POMDP inherits structure from its MDP analog.
2. The M-POMDP inherits structure with respect to the belief distribution from the general POMDP.
3. The M-POMDP yields specialized, tractable solution procedures due to its passive learning environment.

These properties were facilitated by a *separability* assumption on the transition dynamics, akin to the separation principle in stochastic optimal control. One particularly interesting line of research is to consider how M-POMDPs might be used as tractable *approximations* to more general POMDPs in which the separability assumption does not hold. That is, how good is an optimal policy generated from an M-POMDP, for an analogous POMDP in which the separability assumption *almost* holds?

Future research might additionally consider applications of the M-POMDP. One particularly interesting potential application area is healthcare operations under real-time patient

health data. In such applications, a healthcare operator must make decisions (*e.g.* ordering, production, distribution decisions) while simultaneously considering real-time patient health data. This setting would be particularly important for critically ill patients, and personalized gene therapies, in which the patient's health evolves independently of operating decisions, but affects the operator's objectives to both minimize costs and maximize patient well-being.

5.2 The Value of Information and Supply Chain Agility in Managing Uncertainty in Inventory Systems

In Chapter 3, we consider the strategic demand management question: should capital be allocated towards (1) a better information infrastructure (data, forecasts, quantitative talent, *etc*.), or (2) a more agile product architecture and supply chain design in order to more quickly respond to changing demand? We recognize that answering this question requires considering how to analyze and quantify the effects of changes in agility and information quality on optimal decision-making within an operational context. We show how to address this question in a specific inventory control context, but future research might alternatively consider different and/or more complex contexts.

Additionally, we present a notion of information quality in terms of Markov noisy channels, and show that, in this sense, better information quality results in better system performance under an optimal policy. Although mathematically appealing, this notion of information quality is difficult to interpret in practice because it is a *relative* notion of information quality. We can only determine a partial order on the quality of information mechanisms through this notion. However, for a specific numerical example, we were able to present results with respect to an *absolute* measure of information quality. Future research might seek to generalize these results.

5.3 Generating Trust in Development Processes Using Robust, Data-driven Markov Games: An Application to PRESTIGE

In Chapter 4, we consider a two-agent decision-making environment in which a defender must make decisions in a development process, in which there is potential for adversarial manipulation. In this context, we model agent dynamics using a partially observable Markov game (POMG), and we show how to generate a policy that the defender can *trust* — in the sense that the defender policy is robust to an array of attacker policies (potentially due to misapprehension of adversarial objectives, level of rationality, *etc*.) — by combining a generative mechanism for constructing plausible attacker policies and then solving a robust dynamic program that optimizes over the expected objectives with respect to an *a priori* weight distribution, λ.

Traditional reinforcement learning has its mathematical foundations in Markov decision processes (MDPs), and has been applied successfully (and famously) to zero-sum games, such as chess and Go ([44]). However, multi-agent reinforcement learning settings in which one agent does not know the other agent's objectives have not been well-considered, to the knowledge of this author, despite there being many applications (as we discuss in Chapter 4). A future research direction would be to consider how our approach might be adapted to a reinforcement learning setting.

Moreover, there is a development in the artificial intelligence (AI) community towards *explainable AI*. As we discuss in the conclusions of Chapter 4, the weight distribution λ in our robust dynamic program might act as a way to explain the defender's policy behavior in terms of the weights that are placed upon potential adversarial policies. Future research methods might allow the DM to adjust these weights over time, as a foundation for constructing policies that are both *robust and explainable*.

Finally, another future research direction might focus on improving the generative mechanism that we present, using, for example, heuristic or evolutionary search mecha-

nisms such as genetic algorithms.

Appendices

APPENDIX A
SEQUENTIAL DECISION-MAKING AFFECTED BY PARTIALLY OBSERVED EXOGENOUS FORCES

A.1 $L^\natural$-convexity

We now examine a structure from discrete convex analysis, $L^\natural$-convexity, which can be useful in many operations research applications. [64] proved the existence of a monotone optimal policy and a $L^\natural$-convex value function for a completely non-perishable inventory problem which is modeled as a MDP. In this subsection, we show how the results in Section 2.4 can be used to demonstrate that the M-POMDP inherits this optimal policy structure.

We provide a brief introduction to this topic, and refer the reader to [32] for a thorough treatment. Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{Z}$ and e be the ones vector. We begin with a definition of $L^\natural$-convexity and a proposition that demonstrates that $L^\natural$-convexity is a C3 property.

Definition 8. *($L^\natural$-convexity) A real-valued function $g(y)$ defined on an $L^\natural$-convex set $Y \subseteq \mathbb{F}^n$, i.e.*

$$\forall y, y' \in Y, \forall \alpha \in \mathbb{F}^+, \quad y \vee (y - \alpha e) \in Y \text{ and } (y + \alpha e) \wedge y' \in Y$$

is an $L^\natural$-convex function if the function $\psi(y, \xi) = g(y - \xi e), \xi \leq 0$ is subadditive on $Y \times \mathbb{F}^-$.

Proposition 23. *$L^\natural$-convexity is a C3 property.*

The following proposition demonstrates the inheritance of the $L^\natural$-convex value function and an optimal monotone policy function for the M-POMDP, given that the MDP has these properties.

Proposition 24. *Suppose B(b) holds for $\tilde{F}$ the space of real-valued functions on $S \times A$ that are $L^\natural$-convex on S. Then there exists an optimal value-policy function pair $(v^*, \pi^*) \in V \times \Pi$*

such that $v^*(s,x)$ *is* $L^\natural$*-convex in* s *on* S *for all* $x \in X$*, and* $\pi^*(s,x)$ *is non-decreasing in* s *on* S *for all* $x \in X$.

In [64] this structure is derived for a completely-observed lost sales inventory problem, in which the author notes that the monotone optimal policy structure extends under a completely observed Markov-modulated demand process. We note that this is simply a special case of a M-POMDP in which the modulation process (corresponding to the "world" process in [64]) is completely observed.

We note, additionally, that the closely related concept of multi-modularity is also a C3 property since it is mathematically equivalent to $L^\natural$-convexity via a coordinate transformation. As with $L^\natural$-convexity, multi-modularity is useful in certain inventory control problem settings, as in [28].

A.2 Proofs

Proof of Proposition 2. Note that $v(\cdot,x) \in \tilde{V}$ for all $x \in X$ implies that $v(\cdot,\lambda(z',x)) \in \tilde{V}$ for all $(z',x) \in Z \times X$. Recall that

$$Hv(s,x) = \min_{a \in A(s)} \sum_{z'} \sigma(z'|x) h_{z'}\big(s,a,v(\cdot,\lambda(z',x))\big).$$

We know from [46] that single-point properties are preserved under maximization. Let $\mathscr{G}'$ be the set of all real-valued functions on S that possess a C3 property, $\mathscr{P}'$, and let $\mathscr{G} = \{-v : v \in \mathscr{G}'\}$. Then $\mathscr{G}$ also possesses a C3 property, $\mathscr{P}$ (not necessarily $\mathscr{P}'$).

Now, observe that $\sum_{z'} \sigma(z'|x) h_{z'}\big(\cdot,a,v(\cdot,\lambda(z',x))\big) \in \tilde{V}$ since C3 properties are closed under convex combinations. Further, single-point properties are preserved under maximization, so since

$$-Hv(s,x) = \max_{a \in A} \left\{ -\sum_{z'} \sigma(z'|x) h_{z'}\big(s,a,v(\cdot,\lambda(z',x))\big) \right\}$$

we conclude that $Hv(\cdot, x) \in \tilde{V}$ for all $x \in X$. The result follows by applying Corollary 1. □

Proof of Proposition 3. Note that $v(\cdot, x) \in \tilde{V}$ for all $x \in X$ implies that $v(\cdot, \lambda(z', x)) \in \tilde{V}$ for all $(z', x) \in Z \times X$. Recall that

$$Hv(s, x) = \min_{a \in A(s)} \sum_{z'} \sigma(z'|x) h_{z'}\big(s, a, v(\cdot, \lambda(z', x))\big).$$

Since $v(\cdot, x) \in \tilde{V}$ for all $x \in X$, $h_{z'}\big(s, a, v(\cdot, \lambda(z', x))\big)$ satisfies a joint extension of $\mathscr{P}$ on $S \times A$, $\mathscr{P}^*$, by the proposition assumption, for all $(z', x) \in Z \times X$. This implies $\sum_{z'} \sigma(z'|x) h_{z'}\big(s, a, v(\cdot, \lambda(z', x))\big)$ has property $\mathscr{P}^*$ on $S \times A$, as well. The result follows from Proposition 4 in [46]. □

Proof of Proposition 4. Suppose $\bar{v} \in \tilde{V}$. The proof proceeds by first showing $h_{z'}(s, a, \bar{v})$ satisfies $\mathscr{P}^*$ and then applying Proposition 3. Recall

$$h_{z'}(s, a, \bar{v}) = c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a) \bar{v}(s').$$

Note that c satisfies $\mathscr{P}^*$, for all $z' \in Z$ due to (i). Then Proposition 2 of [46] and (ii) guarantee that $\sum_{s'} p(s'|z', s, a) \bar{v}(s')$ has property $\mathscr{P}^*$ for all $z' \in Z$. Since $\mathscr{P}^*$ is a convex cone, $h_{z'}\big(s, a, \bar{v}\big)$ has property $\mathscr{P}^*$ in (s, a) on $S \times A$, for all $z' \in Z$. The result follows by applying Proposition 3. □

Proof of Proposition 6. We proceed by demonstrating that P(b) and P(c) hold and then applying Proposition 1. Suppose $v(\cdot, x) \in \tilde{V}$ for all $x \in X$. Recall, we have

$$Hv(s, x) = \min_{a \in A(s)} \sum_{z'} \sigma(z'|x) h_{z'}\big(s, a, v(\cdot, \lambda(z', x))\big).$$

By B(a), we have that $h_{z'}\big(\cdot,\cdot,v(\cdot,\lambda(z',x))\big) \in \tilde{F}$ for all $(z',x) \in Z \times X$. Further,

$$\sum_{z'} \sigma(z'|x) h_{z'}\big(\cdot,\cdot,v(\cdot,\lambda(z',x))\big) \in \tilde{F}$$

as well, since $\tilde{F}$ is a space of functions possessing a joint C3 property, $\mathscr{P}^*$. By B(b), minimizing over feasible policies from S to A maps functions in $\tilde{F}$ into $\tilde{V}$. We conclude that $Hv(\cdot,x) \in \tilde{V}$ for all $x \in X$ and P(b) holds.

By the same logic, since $\sum_{z'} \sigma(z'|x) h_{z'}\big(\cdot,\cdot,v(\cdot,\lambda(z',x))\big) \in \tilde{F}$, B(c) guarantees that P(c) holds as well. The conclusion follows by Proposition 1. □

Proof of Proposition 7. Suppose $v_n(s,\cdot)$ is piecewise linear and concave in x. It suffices to show $v_{n+1}(s,\cdot) = Hv_n(s,\cdot)$ is piecewise linear and concave in x.

If $v_n(s,\cdot)$ is piecewise linear and concave, then that implies there exists a vector on μ, $\vec{\alpha}(s)$, such that $v_n(s,\cdot)$ can be expressed as follows.

$$\begin{aligned} v_n\big(s,\lambda(\cdot|z',x)\big) &= v_n\left(s,\frac{\sum_{\mu\in M} x(\mu)P[z',\cdot|u]}{\sum_{u\in U} x(\mu)P[z'|\mu]}\right) \\ &= v_n\left(s,\frac{\sum_{\mu\in M} x(u)P[z',\cdot|\mu]}{\sigma(z'|x)}\right) \\ &= \left[\frac{\sum_{\mu\in M} x(\mu)P[z',\cdot|\mu]}{\sigma(z'|x)}\right]^T \vec{\alpha}(s) \end{aligned}$$

Now, we plug into the value iteration equation, $v_{n+1} = Hv_n$.

$$
\begin{aligned}
v_{n+1}(s,x) &= \min_{a\in A(s)} \left\{ c(s,x,a) + \beta \sum_{s'\in S} \sum_{z'\in Z} \sigma(z'|x) p(s'|z',s,a) v_n\big(s', \lambda(\cdot|z',x)\big) \right\} \\
&= \min_{a\in A(s)} \Bigg\{ c(s,x,a) \\
&\quad + \beta \sum_{s'\in S} \sum_{z'\in Z} \sigma(z'|x) p(s'|z',s,a) \left[\frac{\sum_{\mu\in M} x(\mu) P[z',\cdot|\mu]}{\sigma(z'|x)} \right]^T \vec{\alpha}_{(z',x)}(s) \Bigg\} \\
&= \min_{a\in A(s)} \left\{ x^T c(s,\cdot,a) + \beta \sum_{z',s'} p(s'|z',s,a) \big[P(z')x\big] \vec{\alpha}_{(z',x)}(s) \right\} \\
&= x^T \left[c(s,\cdot,a^*) + \beta \sum_{z',s'} p(s'|z',s,a^*) P(z') \vec{\alpha}_{(z',x)}(s) \right] \\
&= x^T \vec{\alpha}'(s)
\end{aligned}
$$

So successive approximations preserve piecewise linearity and concavity of the value functions. Concavity is preserved in the limit as $n \to \infty$. Piecewise linearity is not in general preserved in the limit, but may be under some specialized conditions, such as the finite transience condition in [48]. □

Proof of Proposition 8. Proof follows along the lines of Proposition 1 in [29], by induction on the value iteration iterates. Suppose $v_0 = 0$ and that $v_n(s,x)$ is nondecreasing in x on X, in the sense of monotone likelihood ratio, for all $s \in S$. Suppose $x \succeq_{LR} \bar{x}$.

We first show that σ preserves MLR-order and that $\mathbb{E}c$ is non-decreasing in z' on Z.

$$
\begin{aligned}
x \geq_{LR} \bar{x} \quad &\iff \quad x \geq_S \bar{x} \\
&\iff \quad \sum_{\mu \geq q} x(\mu) f(\mu) \geq \sum_{\mu \geq q} \bar{x}(\mu) f(\mu), \quad \forall q \in M, f \text{ nondecreasing}
\end{aligned}
$$

by Lemma 1.1 in [29]

$$
\implies \quad \sum_{\mu \geq q} x(\mu) \sum_{\mu'} P[z', \mu' | \mu] \geq \sum_{\mu \geq q} \bar{x}(\mu) \sum_{\mu'} P[z', \mu' | \mu], \quad \forall q \in M, z' \in Z
$$

by (d)

$$
\begin{aligned}
&\implies \quad \sigma(z'|x) \geq \sigma(z'|\bar{x}), \quad \forall z' \in Z \\
&\implies \quad \sigma(\cdot|x) \geq_S \sigma(\cdot|\bar{x}) \\
&\implies \quad \sum_{z'} \sigma(z'|x) c(s, z', a) \geq \sum_{z'} \sigma(z'|\bar{x}) c(s, z', a), \quad \forall (s, a) \in S \times A,
\end{aligned}
$$

by condition (a) and Lemma 1.1 in [29].

Next, we show that λ preserves MLR-order.

$$
\text{condition (c) and } x \geq_{LR} \bar{x} \implies \quad \sum_{\mu} x(\mu) P[z', \cdot | \mu] \geq_{LR} \sum_{\mu} \bar{x}(\mu) P[z', \cdot | \mu], \quad \forall z' \in Z
$$

by Lemma 1.3.1 in [29]

$$
\implies \quad \lambda(z', x) \geq_{LR} \lambda(z', \bar{x}), \quad \forall z' \in Z
$$

by Lemma 1.2.2 in [29].

Finally, we want to show λ is non-decreasing in z', in the sense of MLR. Suppose $z' \geq \bar{z}$,

$\mu' \geq \bar{\mu}$.

$$\text{condition (e)} \implies \sum_{\mu} x(\mu) P[z', \mu'|\mu] \cdot \sum_{\mu} x(\mu) P[\bar{z}, \bar{\mu}|\mu] \geq \sum_{\mu} x(\mu) P[z', \bar{\mu}|\mu] \cdot \sum_{\mu} x(\mu) P[\bar{z}, \mu'|\mu]$$

by Lemma 1.3.2 in [29]

$$\iff \lambda(\mu'|z', x) \cdot \lambda(\bar{\mu}|\bar{z}, x) \geq \lambda(\bar{\mu}|z', x) \cdot \lambda(\mu'|\bar{z}, x)$$

$$\iff \lambda(z', x) \geq_{LR} \lambda(\bar{z}, x).$$

Altogether, we show that $v_{n+1}(s, x) \geq v_{n+1}(s, \bar{x})$.

$$\begin{aligned}
h_{z'}\big(s, a, v_n(\cdot, \lambda(z', x))\big) &= c(s, z', a) + \beta \sum_{s'} p(s'|z', s, a) v_n\big(s', \lambda(z', x)\big) \\
&\geq c(s, \bar{z}, a) + \beta \sum_{s'} p(s'|\bar{z}, s, a) v_n\big(s', \lambda(\bar{z}, x)\big) \\
&\quad \text{by induction hypothesis, condition (b), and } \lambda(z', x) \geq_{LR} \lambda(\bar{z}, x) \\
&\geq c(s, \bar{z}, a) + \beta \sum_{s'} p(s'|\bar{z}, s, a) v_n\big(s', \lambda(\bar{z}, \bar{x})\big) \\
&\quad \text{by induction hypothesis and } \lambda(\bar{z}, x) \geq_{LR} \lambda(\bar{z}, \bar{x}) \\
&= h_{\bar{z}}\big(s, a, v_n(\cdot, \lambda(\bar{z}, \bar{x}))\big).
\end{aligned}$$

Since this holds for all a, and $\sigma(\cdot|x) \geq_S \sigma(\cdot|\bar{x})$, we conclude by Lemma 1.1 in [29] that $v_{n+1}(s, x) \geq v_{n+1}(s, \bar{x})$. Monotonicity is preserved in the limit as $n \to \infty$, so $v^*(s, x) \geq v^*(s, \bar{x})$ for all $s \in S$. □

Proof of Proposition 9. Note that

$$\sigma(z'|x) = \sum_{\tilde{z}} \xi(z'|\tilde{z}) \tilde{\sigma}(\tilde{z}|x)$$

$$\lambda(\mu'|z', x) = \Lambda(z'|\tilde{z}, x) \tilde{\lambda}(\mu'|\tilde{z}, x),$$

where $\Lambda(z'|\tilde{z}, x) = \frac{\xi(z'|\tilde{z}) \tilde{\sigma}(\tilde{z}|x)}{\sigma(z'|x)}$, which are convex multipliers since $\sum_{\tilde{z}} \Lambda(z'|\tilde{z}, x) = 1$.

Note that Proposition 7 and Jensen's inequality imply

$$\sum_{\tilde{z}} \Lambda(z'|\tilde{z},x)\tilde{v}(s',\tilde{\lambda}(\tilde{z},x)) \le \tilde{v}(s',\lambda(\tilde{z},x)). \tag{A.1}$$

Proof is by induction on the value iteration iterates. The base case $\tilde{v}_0 \le v_0$ is satisfied by assumption. Suppose, for induction, that $\tilde{v}_n \le v_n$, and that $v_{n+1} = Hv_n$ and $\tilde{v}_{n+1} = \tilde{H}\tilde{v}_n$.

$$\begin{aligned}
v_{n+1}(s,x) &= \min_{a\in A(s)} \sum_{z'} \sigma(z'|x)\left[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) v_n\big(s',\lambda(z',x)\big)\right] \\
&\ge \min_{a\in A(s)} \sum_{z'} \sigma(z'|x)\left[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) \tilde{v}_n\big(s',\lambda(z',x)\big)\right] \\
&\quad \text{by the induction hypothesis} \\
&\ge \min_{a\in A(s)} \sum_{z'} \sigma(z'|x)\left[c(s,z',a) + \beta \sum_{s'}\sum_{\tilde{z}} p(s'|z',s,a)\Lambda(z'|\tilde{z},x) \tilde{v}_n\big(s',\tilde{\lambda}(\tilde{z},x)\big)\right] \\
&\quad \text{by Equation (A.1)} \\
&= \min_{a\in A(s)} \sum_{\tilde{z}}\sum_{z'} \xi(z'|\tilde{z})\tilde{\sigma}(\tilde{z}|x)\left[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) \tilde{v}_n\big(s',\tilde{\lambda}(\tilde{z},x)\big)\right] \\
&\quad \text{by plugging in for } \sigma \text{ and } \Lambda \\
&= \min_{a\in A(s)} \sum_{\tilde{z}} \tilde{\sigma}(\tilde{z}|x)\left[\sum_{z'}\xi(z'|\tilde{z})c(s,z',a) + \beta \sum_{s'}\sum_{z'} \xi(z'|\tilde{z}) p(s'|z',s,a) \tilde{v}_n\big(s',\tilde{\lambda}(\tilde{z},x)\big)\right] \\
&= \min_{a\in A(s)} \sum_{\tilde{z}} \tilde{\sigma}(\tilde{z}|x)\left[\tilde{c}(s,\tilde{z},a) + \beta \sum_{s'} \tilde{p}(s'|\tilde{z},s,a) \tilde{v}_n\big(s',\tilde{\lambda}(\tilde{z},x)\big)\right] \\
&\quad \text{by proposition conditions (a) and (b).} \\
&= \tilde{v}_{n+1}(s,x)
\end{aligned}$$

The result follows by taking the limit as $n \to \infty$. □

Proof of Proposition 11. An equivalent representation of $\mathscr{P}^*_{\tilde{V}}$ for all $v \in V$ such that $v(\cdot,x) \in \tilde{V}$ for all $x \in X$, is

$$\sum_{s'} v\big(s',\lambda(z',x)\big) \sum_{j\in J_k} \gamma_j p(s'|z',s_j,a_j) \le \sum_{s'} v\big(s',\lambda(z',x)\big) \sum_{i\in I_k} \gamma_i p(s'|z',s_i,a_i), \quad \forall k \in K.$$

Since state dynamics can be described functionally, we may rewrite the above inequality by applying f and interchanging summations.

$$\sum_{j\in J_k} \gamma_j v\big(f(z', s_j, a_j), \lambda(z', x)\big) \leq \sum_{i\in I_k} \gamma_i v\big(f(z', s_i, a_i), \lambda(z', x)\big), \quad \forall k \in K$$

Note that this is the inequality test of satisfaction description of $\mathscr{P}^*$. □

Proof of Proposition 12. By the inequality test of satisfaction, monotonicity, and Jensen's inequality, for all $k \in K$ and $(z', x) \in Z \times X$,

$$\begin{aligned} f(z', s_k, a_k) &\leq_S \sum_{j\in J_k} \gamma_j f(z', s_j, a_j) \\ v\big(f(z', s_k, a_k), \lambda(z', x)\big) &\leq v\left(\sum_{j\in J_k} \gamma_j f(z', s_j, a_j), \lambda(z', x)\right) \\ &\leq \sum_{j\in J_k} \gamma_j v\big(f(z', s_j, a_j), \lambda(z', x)\big). \end{aligned}$$

□

The following two lemmas will be useful in our proof of Proposition 13 and is based upon Lemmas 4.7.2 and 4.7.1 in [37], respectively, adapted to our setting here.

Lemma 1. *Let y, y' be non-negative real-valued functions on S such that*

$$\sum_{s\geq k} y(s) \leq \sum_{s\geq k} y'(s), \text{ for all } k \in S,$$

and $\sum_{s\in S} y(s) = \sum_{s\in S} y'(s)$. Suppose $\tilde{V}$ is the space of all non-increasing real-valued functions on S. Then $\tilde{v} \in \tilde{V}$ implies

$$\sum_{s\in S} y(s)\tilde{v}(s) \leq \sum_{s\in S} y'(s)\tilde{v}(s).$$

Lemma 2. *If $f : S \times A \mapsto \mathbb{R}$ is subadditive, then $\pi \in \Pi$ such that*

$$\pi(s) = \min\left\{a' : a' \in \arg\min_{a \in A} f(s,a)\right\}$$

is non-increasing in s on S.

Proof of Proposition 13. We begin by naturally defining our structured spaces. Let

$\tilde{V} \triangleq \{\bar{v} : \bar{v}$ non-increasing in s on $S\}$

$\tilde{\Pi} \triangleq \{\bar{\pi} : \bar{\pi}$ non-decreasing in s on $S\}$

$\tilde{P} \triangleq \{\bar{p} : \bar{p}$ non-increasing in s on S for all $a \in A$ and

subadditive on $S \times A$ in the sense of first-order stochastic dominance$\}$

$\tilde{C} \triangleq \{\bar{c} : \bar{c}$ non-increasing in s on S for all $a \in A$, subadditive on $S \times A\}$

$\tilde{F} \triangleq \{f : f$ is subadditive on $S \times A$ and non-increasing in s on S for all $a \in A.\}$

We want to prove that (i) - (iv) imply P(a) and B(a) - B(c), at which point we will apply Proposition 6 to get the desired result. First, P(a) holds because $\tilde{V}$ is the set of non-increasing value functions on S, a closed space.

We aim to show B(a) holds. Suppose $\tilde{v} \in \tilde{V}$. By (iv), we have for $s^+, s^- \in S$ and $a^+, a^- \in A(s^+)$ such that $s^+ \geq s^-$ and $a^+ \geq a^-$,

$$\sum_{s' \geq k} \left[p(s'|z', s^+, a^+) + p(s'|z', s^-, a^-)\right] \leq \sum_{s' \geq k} \left[p(s'|z', s^+, a^-) + p(s'|z', s^-, a^+)\right], \forall k \in S, z' \in Z.$$

By Lemma 1, $\tilde{v} \in \tilde{V}$ implies

$$\sum_{s' \geq k} \left[p(s'|z', s^+, a^+) + p(s'|z', s^-, a^-)\right]\tilde{v}(s') \leq \sum_{s' \geq k} \left[p(s'|z', s^+, a^-) + p(s'|z', s^-, a^+)\right]\tilde{v}(s'),$$

for all $k \in S$, $z' \in Z$. We conclude the $\mathbb{E}[\tilde{v}|z', s, a]$ is subadditive in (s, a) on $S \times A$ for all $z' \in Z$. It follows from (iii) and the fact that subadditivity is a C3 property that $h_{z'}(s, a, \tilde{v})$

is subadditive in (s, a) on $S \times A$ for all $z' \in Z$. It remains to show that $h_{z'}$ is non-increasing in s on S, which follows from (i), (ii), and Lemma 1. So B(a) holds.

Now consider B(b). Suppose $f \in \tilde{F}$. B(b) holds by the following simple argument:

$$\begin{aligned}\min_{a \in A(s)} f(s,a) &= f(s, a_s^*) \\ &\geq f(s', a_s^*) \qquad \text{(by } f \text{ nonincreasing on } S\text{)} \\ &\geq \min_{a \in A(s')} f(s', a).\end{aligned}$$

B(c) holds by Lemma 2. The conclusion follows from Proposition 6. □

Proof of Proposition 23. We show first that $L^\natural$-convexity is a convex cone and then show that it is closed under the topology of pointwise convergence.

Suppose g, h are $L^\natural$-convex functions defined on an $L^\natural$-convex set, X, with corresponding functions $\psi_g(x, \xi) = g(x - \xi e)$ and $\psi_h(x, \xi) = h(x - \xi e)$, and $\alpha, \beta \geq 0$.

Since g, h are $L^\natural$-convex functions, we know that ψ_g and ψ_h are subadditive on $X \times \mathbb{F}^-$, which gives us the following inequalities, where $x_1 \leq x_2$ and $\xi_1 \leq \xi_2$,

$$\begin{aligned}\psi_g(x_1, \xi_1) + \psi_g(x_2, \xi_2) &\leq \psi_g(x_1, \xi_2) + \psi_g(x_2, \xi_1) \\ \psi_h(x_1, \xi_1) + \psi_h(x_2, \xi_2) &\leq \psi_h(x_1, \xi_2) + \psi_h(x_2, \xi_1).\end{aligned}$$

Taking the conic combination of these inequalities with weights α, β, and defining $\psi_f = \alpha\psi_g + \beta\psi_h$ and $f = \alpha g + \beta h$, we get

$$\psi_f(x_1, \xi_1) + \psi_f(x_2, \xi_2) \leq \psi_f(x_1, \xi_2) + \psi_f(x_2, \xi_1).$$

We conclude that ψ_f is subadditive on $X \times \mathbb{F}^-$, $f = \alpha g + \beta h$ is $L^\natural$-convex, and $L^\natural$-convexity is a convex cone. It remains to show that $L^\natural$-convexity is closed under the topology of pointwise convergence, which follows from the fact that subadditive is closed under the

topology of pointwise convergence. □

Proof of Proposition 24. Let $\tilde{V} = \{\bar{v} \in \bar{V} : \bar{v} \text{ is } L^\natural\text{-convex on } S\}$. We show that B(a) - B(c) hold, and then apply Proposition 6. First, P(a) holds from Proposition 23, since $L^\natural$-convexity is a C3 property. B(a) holds by the proposition assumption. B(b) and B(c) hold by Lemmas 2 and 3, respectively, in [64]. □

Proof of Proposition 14. We first show that separability is a convex cone. Suppose we have two separable functions, f and g, which map $S \times A$ to $\mathbb{R}$, and conic weights $\alpha, \beta \geq 0$. Clearly,

$$\alpha f(s,a) + \beta g(s,a) = \alpha K_f(a) + \beta K_g(a) + \alpha L_f(s) + \beta L_g(s)$$

so that $\alpha f + \beta g$ is a separable function and hence that separability is a convex cone. It remains to show that f is closed under the topology of pointwise convergence.

Suppose we have a sequence of separable functions $\{f_n\}$ such that f_n converges pointwise to $f = \lim_{n\to\infty} f_n$. Note that $f(s,a) = \lim_{n\to\infty} f_n(s,a) = \lim_{n\to\infty} [L_n(s) + K_n(a)]$. We conclude, by linearity of limits, that $f_n(s,a) = L_n(s) + K_n(a) \to L(s) + K(a) = f(s,a)$ as $n \to \infty$, where $L(s) = \lim_{n\to\infty} L_n(s)$ and $K(a) = \lim_{n\to\infty} K_n(a)$. □

Proof of Proposition 15. Suppose $v(\cdot, x) \in \tilde{V}$ for all $x \in X$. We begin by defining our structured spaces:

$$\tilde{\Pi} \triangleq \{\tilde{\pi} : \exists a \in A : \bar{\pi}(s) = a, \forall s \in S\}$$

$$\tilde{V} \triangleq \{\tilde{v} : \exists L \in \mathscr{L} : \bar{v}(s) = L(s)\}$$

$$\tilde{C} \triangleq \{\tilde{c} : \exists K \in \mathscr{K}, L \in \mathscr{L} : \bar{c}(s,a) = K(a) + L(s)\}$$

$$\tilde{P} \triangleq \{\tilde{p} : \bar{p}(\cdot|s,a) = \bar{p}(\cdot|a)\}$$

$$\tilde{F} \triangleq \{f : \exists K \in \mathscr{K}, L \in \mathscr{L} : \bar{f}(s,a) = K(a) + L(s)\}.$$

We want to show that there exists a set $\{a(x), x \in X\}$ such that $\pi^*(s,x) = a(x)$ for all $(s,x) \in S \times X$ is stationary optimal by showing that P(a), B(a), B(b), and B(c) hold.

P(a) holds trivially. We aim to show B(a) holds. Suppose $\tilde{v} \in \tilde{V}$. Observe that (i) and (ii) are equivalent to $p(\cdot|z',\cdot,\cdot) \in \tilde{P}$ for all $z' \in Z$ and $c(\cdot,z',\cdot) \in \tilde{C}$ for all $z' \in Z$, which imply that

$$\begin{aligned} h_{z'}(s,a,\tilde{v}) &= c(s,z',a) + \beta \sum_{s' \in S} p(s'|z',s,a)\tilde{v}(s') \\ &= K(z',a) + L(s,z') + \beta \sum_{s' \in S} p(s'|z',a)L(s') \in \tilde{F}, \text{ for all } z' \in Z. \end{aligned}$$

B(b) trivially holds. Further, separable functions when maximized yield state-invariant optimal policies (maximizing $L(s)+K(a)$ over a is equivalent to maximizing $K(a)$ over a for all s). So B(c) holds. By Proposition 6 we conclude that there exists a set $\{a(x), x \in X\}$ such that $\pi^*(s,x) = a(x)$ for all $(s,x) \in S \times X$ is stationary optimal.

It remains to show that $\pi^*(s,x) = a^*(x)$ for all $s \in S$, the myopic minimizer of the function $G(x,a)$. An inductive argument, which follows along the lines of the proof given in [47] proves this result.

Let $L(s,x) = \mathbb{E}[L(s,z')|x]$ and $K(x,a) = \mathbb{E}[K(z',a)|x]$. The value function of the M-POMDP, under any policy π is defined as follows, where $x_{t+1} = \lambda(z_{t+1},x_t)$ and $a_t = \pi(s_t,x_t)$:

$$v^{\pi}(s_0,x_0) = E\left[\sum_{t=0}^{\infty} \beta^t c(s_t,z_{t+1},a_t)|s_0,x_0\right] \tag{A.2}$$

$$= E\left[\sum_{t=0}^{\infty} \beta^t \left[K(x_t,a_t) + L(s_t,x_t)\right]|s_0,x_0\right], \tag{A.3}$$

and where (3) follows from application of assumption (a). From assumption (b), $s_{t+1} \sim$

$\gamma(a_t, z_{t+1})$, where γ is a random variable depending only on a_t and z_{t+1}. Then,

$$
\begin{aligned}
v^{\pi}(s_0, x_0) &= \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \left[K(x_t, a_t) + L(s_t, x_t) \right] | s_0, x_0 \Bigg] \\
&= K(x_0, a_0) + L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=1}^{\infty} \beta^t \big[K(x_t, a_t) + L\big(\gamma(a_{t-1}, z_t), x_t\big) \big] | s_0, x_0 \Bigg] \\
&= L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \big[K(x_t, a_t) + \beta L\big(\gamma(a_t, z_{t+1}), x_{t+1}\big) \big] | s_0, x_0 \Bigg] \\
&= L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \big[K(x_t, a_t) + \beta L\big(\gamma(a_t, z_{t+1}), \lambda(z_{t+1}, x_t)\big) \big] | s_0, x_0 \Bigg] \\
&= L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \Bigg[K(x_t, a_t) \\
&+ \beta \sum_{z''} \sigma\big(z'' | \lambda(z_{t+1}, x_t)\big) \sum_{s'} p(s' | z_{t+1}, a_t) L(s', z'') \Bigg] | s_0, x_0 \Bigg] \\
&= L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t G(x_t, a_t) | s_0, x_0 \Bigg] \\
&\geq L(s_0, x_0) + \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t G(x_t, a^*(x_t)) | s_0, x_0 \Bigg].
\end{aligned}
$$

We conclude that the policy $\pi^*(s, x) = a^*(x)$ for all $s \in S$ is stationary and optimal. □

Proof of Proposition 16. We prove the optimal policy result by demonstrating that this inventory problem satisfies the sufficient conditions for a myopic optimal policy presented in Proposition 15. Note, first that the problem can be equivalently transformed to consider the action space to be $A = \mathbb{Z}_{\geq 0}$ the amount of replacement inventory ordered, or as $Y(s) = \mathbb{Z}_{\geq s}$, the order-up-to level, where any order up to level y can be expressed as $y = s + a$. Then, the operator H can be rewritten as

$$
Hv(s, x) = \min_{y \geq s} \sum_{z'} \sigma(z' | x) \left[\hat{c}(z', y) + \beta \sum_{s'} p(s' | z', y) v\left(s', \lambda(z', x)\right) \right].
$$

We note that by replacing the role of a with y in the Proposition 15 insures that c is separable and p is dependent on only z' and y. If we define $y^*(x)$ to be the minimizer of $\sum_{z'} \sigma(z' | x) \hat{c}(y, z')$, then by applying Proposition 15, we infer that the policy (under action

space A) $\pi^*(s,x) = y^*(x) - s$ for all $(s,x) \in S \times X$ is stationary optimal.

Finally, we prove that $v^*(s,x)$ is non-decreasing and convex in s for all $x \in X$. Define $\tilde{V}$ to be $\{v \in V : v \text{ non-decreasing and convex }\}$. We have a functional description of dynamics, so H may be rewritten as

$$Hv(s,x) = \min_{y \geq s} \sum_{z'} \sigma(z'|x) \left[\hat{c}(z',y) + \beta v\Big(f(z',y), \lambda(z',x)\Big)\right].$$

Note that f is non-decreasing and convex in s for all $a \in A$ (through y) and $z' \in Z$, with respect to the usual order on S, a single-point property. Suppose $v(\cdot,x)$ is non-decreasing and convex on S for all $x \in X$. By Proposition 8 we conclude that $v\Big(f(z',\cdot), \lambda(z',x)\Big) \in \tilde{V}$ for all $a \in A$. Due to the separability of $\hat{c}$, we know that $\hat{c}(z',y) + \beta v\Big(f(z',y), \lambda(z',x)\Big)$ is convex in s for all $a \in A$, and conclude that $Hv(\cdot,x) \in \tilde{V}$ for all $x \in X$. Thus P(a) and P(b) hold, so $v^*(\cdot,x) \in \tilde{V}$ for all $x \in X$. □

Proof of Proposition 17. By induction on the value iteration iterates. Let $\tilde{v}_{z',0} = 0$ for all $z' \in Z$ and $v_0 = 0$. Suppose $v_n(s,x) \geq \sum_{z'} \sigma(z'|x)\tilde{v}_{z',n}(s,x)$ for all $(s,x) \in S \times X$, where $v_{n+1} = Hv_n$ and $\tilde{v}_{z',n+1} = \tilde{H}_{z'}\tilde{v}_{z',n}$ for all $z' \in Z$, n.

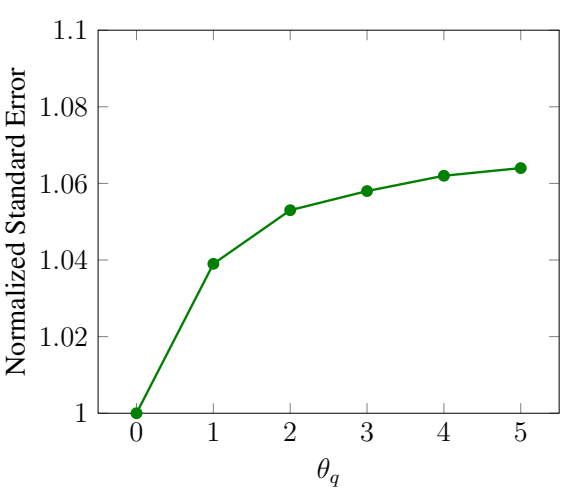
$$\begin{aligned}
v_{n+1}(s,x) &= \min_{a \in A(s)} \sum_{z'} \sigma(z'|x) \left[c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) v_n\Big(s', \lambda(z',x)\Big)\right] \\
&\geq \sum_{z'} \sigma(z'|x) \min_{a \in A(s)} \left\{c(s,z',a) + \beta \sum_{s'} p(s'|z',s,a) v_n\Big(s', \lambda(z',x)\Big)\right\} \\
&\geq \sum_{z'} \sigma(z'|x) \min_{a \in A(s)} \Bigg\{c(s,z',a) \\
&\quad + \beta \sum_{s'} p(s'|z',s,a) \sum_{z''} \sigma\Big(z''|\lambda(z',x)\Big) \tilde{v}_{z'',n}\Big(s', \lambda(z',x)\Big)\Bigg\} \\
&\quad \text{by the induction hypothesis} \\
&= \sum_{z'} \sigma(z'|x) \tilde{v}_{z',n}(s,x).
\end{aligned}$$

The result follows by taking the limit as $n \to \infty$. □

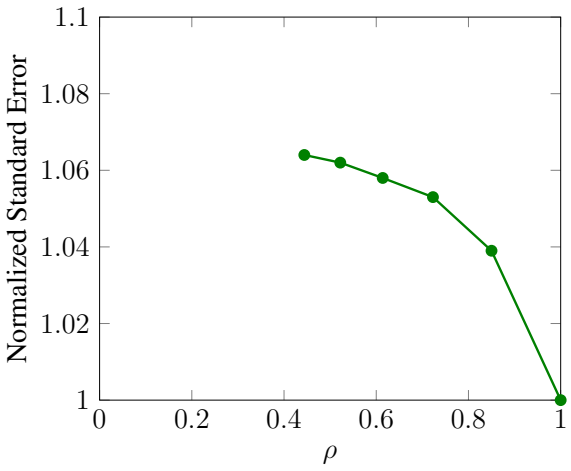
APPENDIX B

THE VALUE OF INFORMATION AND AGILITY IN MANAGING DEMAND UNCERTAINTY

B.1 Figures

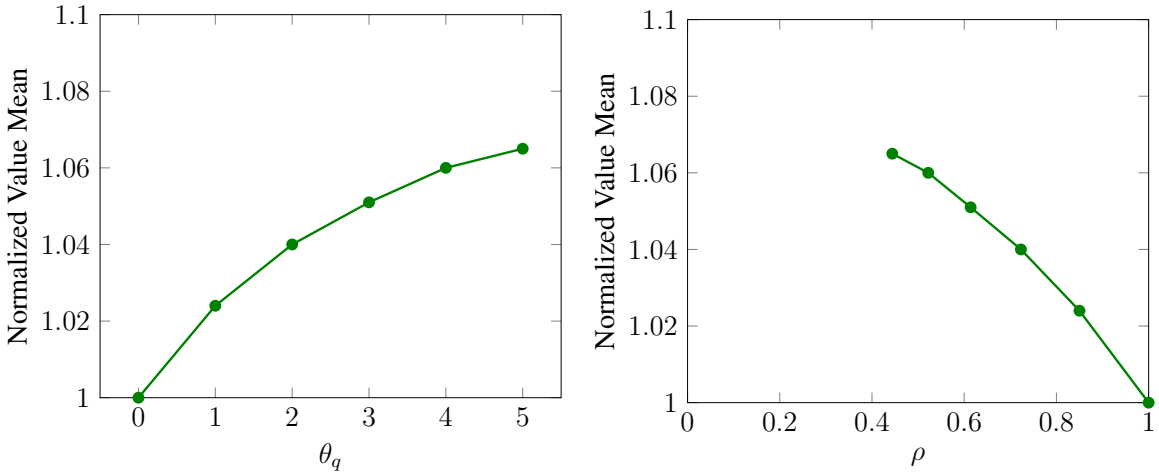


(a) Estimated relative marginal standard error of AOD information, by varying θ_q. Standard errors are normalized to $\theta_q = 0$.



(b) Estimated relative marginal standard error of AOD information, by varying ρ-perfect information quality. Standard errors are normalized to $\rho = 1$.

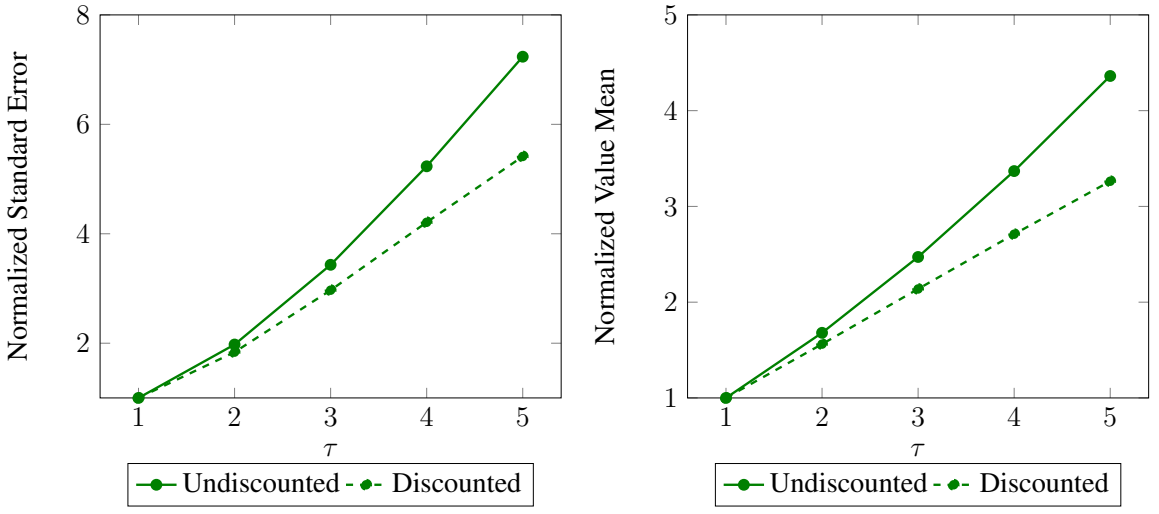
Figure B.1: Marginal effects of AOD information on standard error.



(a) Estimated relative marginal value of AOD information, by varying θ_q. Value means are normalized to $\theta_q = 0$.

(b) Estimated relative marginal value of AOD information, in the ρ-perfect representation. Value means are normalized to $\theta_q = 0$.

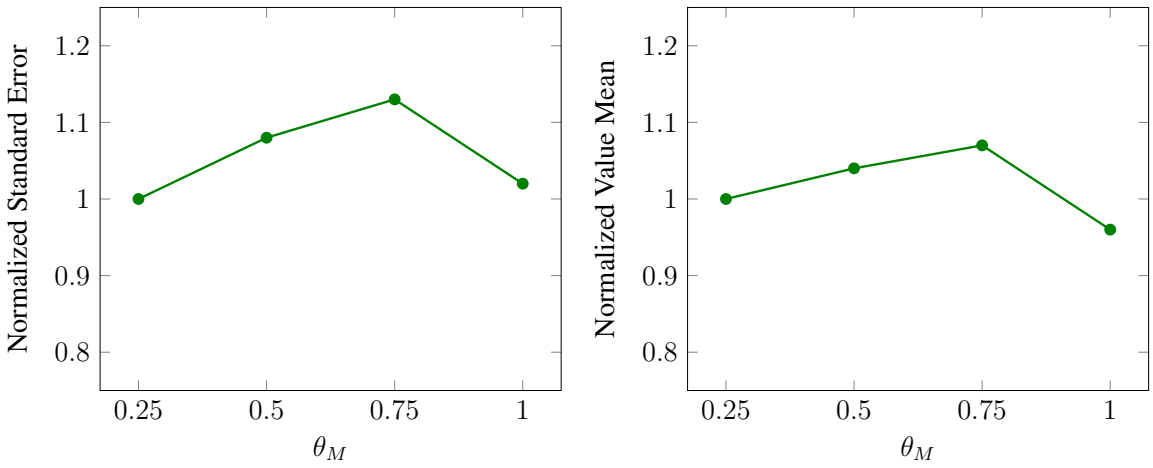
Figure B.2: Marginal value of AOD information and stock-out robustness.



(a) Normalized value mean for various lead times, τ. Discounted standard errors are multiplied by $\beta^{\tau-1}$.

(b) Normalized value mean for various lead times, τ. Discounted value means are multiplied by $\beta^{\tau-1}$.

Figure B.3: Marginal effects of τ.



(a) Normalized standard error for various modulation parameters, θ_M.

(b) Normalized value mean for various modulation parameters, θ_M.

Figure B.4: Marginal effects of θ_M.

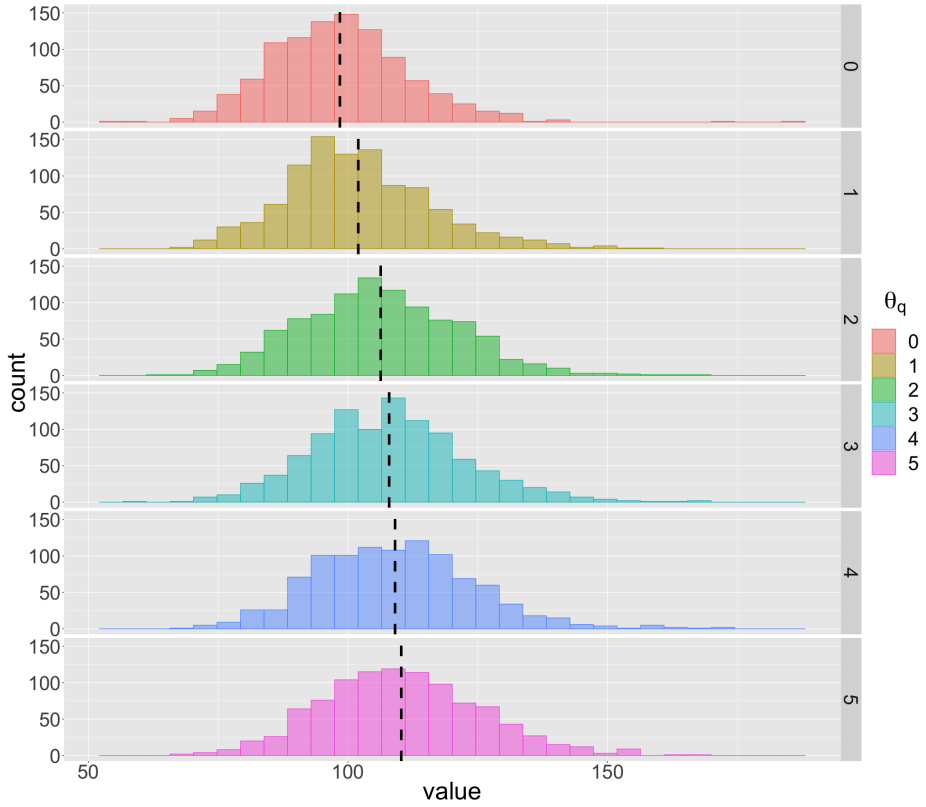


Figure B.5: Histogram of simulated values by θ_q, for fixed $\theta_M = 0.5$, $p = 3$, $\tau = 2$. The dashed lines represent the sample means.

Table B.1: The regression output for the log-linear regressions for value mean, standard error, stock-outs, and attainability violations.

	$\log(\text{SE}_\theta)$		$\log(v_\theta)$		$\log(\text{SO}_\theta)$		$\log(\text{ATN}_\theta)$	
Regressor	t-statistic	b	t-statistic	b	t-statistic	b	t-statistic	b
Intercept	174.26	1.900	634.55	4.058	361.21	8.432	244.87	7.713
$\tau = 1$	–	0	–	0	–	0	–	0
$\tau = 2$	79.00	0.681	102.64	0.519	18.79	0.354	1.77	0.062
$\tau = 3$	143.11	1.233	179.06	0.905	16.93	0.319	4.05	0.143
$\tau = 4$	192.05	1.655	240.28	1.215	14.73	0.278	5.82	0.205
$\tau = 5$	229.65	1.979	291.39	1.473	11.61	0.219	6.61	0.233
$\theta_M = 0.25$	–	0	–	0	–	0	–	0
$\theta_M = 0.50$	9.39	0.072	8.51	0.038	4.31	0.073	23.62	0.744
$\theta_M = 0.75$	16.02	0.124	15.64	0.071	3.78	0.064	41.09	1.294
$\theta_M = 1$	2.72	0.021	-9.58	-0.043	-21.38	-0.361	-5.82	-0.183
$\theta_p = 3$	–	0	–	0	–	–	–	–
$\theta_p = 4$	-3.85	-0.030	12.21	0.055	–	–	–	–
$\theta_p = 5$	-3.80	-0.029	21.76	0.098	–	–	–	–
$\theta_p = 6$	-3.80	-0.029	30.11	0.136	–	–	–	–
θ_p (numeric)	–	–	–	–	-3.22	-0.037	–	–
$\theta_q = 0$	–	0	–	0	–	0	–	–
$\theta_q = 1$	4.04	0.038	4.36	0.024	-4.64	-0.096	–	–
$\theta_q = 2$	5.43	0.051	7.14	0.040	-9.42	-0.194	–	–
$\theta_q = 3$	6.01	0.057	9.04	0.050	-21.13	-0.436	–	–
$\theta_q = 4$	6.36	0.060	10.58	0.059	-24.65	-0.509	–	–
$\theta_q = 5$	6.54	0.062	11.42	0.063	-26.85	-0.555	–	–
Adjusted R^2	99.3%		99.6%		84.4%		85.7%	

B.2 Proofs

Proof of Proposition 18. We use the notation $s_{[t+1:t+\tau]} \triangleq \{s_{t+1}, \ldots, s_{t+\tau}\}$, and likewise defined for the other processes $\{z_t\}$, $\{d_t\}$, $\{\mu_t\}$, and $\{a_t\}$. It is sufficient to show that the following relationship holds:

$$P\Big[s_{[t+1:t+\tau]}, z_{[t+1:t+\tau]}, d_{[t+1:t+\tau]}, \mu_{[t+1:t+\tau]} \big| \mathscr{I}_t\Big] = P\left[s_{[t+1:t+\tau]}, z_{[t+1:t+\tau]}, d_{[t+1:t+\tau]}, \mu_{[t+1:t+\tau]} \big| s_t, d_t, a_{[t-\tau:t-1]}, x_t\right],$$

where $\mathscr{I}_t = \left\{s_{[0:t]}, z_{[1:t]}, d_{[0:t]}, a_{[-\tau:t-1]}, x_0\right\}$. Consider the following:

$$
\begin{aligned}
&P\Big[s_{[t+1:t+\tau+1]}, z_{[t+1:t+\tau+1]}, d_{[t+1:t+\tau+1]}, \mu_{[t+1:t+\tau]}\big|\mathscr{I}_t\Big] \\
&\quad = P\Big[s_{t+\tau}, z_{t+\tau}, d_{t+\tau}, \mu_{t+\tau}\big|s_{[t+1:t+\tau-1]}, z_{[t+1:t+\tau-1]}, d_{[t+1:t+\tau-1]}, \mu_{[t+1:t+\tau-1]}, \mathscr{I}_t\Big]\cdots \\
&\quad\cdots P\Big[s_{[t+1:t+\tau-1]}, z_{[t+1:t+\tau-1]}, d_{[t+1:t+\tau-1]}, \mu_{[t+1:t+\tau-1]}\big|\mathscr{I}_t\Big] \\
&\quad = P\Big[s_{t+\tau}, z_{t+\tau}, d_{t+\tau}, \mu_{t+\tau}\big|s_{t+\tau-1}, d_{t+\tau-1}, \mu_{t+\tau-1}, a_{t-1}\Big]\cdots \\
&\quad\cdots P\Big[s_{[t+1:t+\tau-1]}, z_{[t+1:t+\tau-1]}, d_{[t+1:t+\tau-1]}, \mu_{[t+1:t+\tau-1]}\big|\mathscr{I}_t\Big] \\
&\quad = \prod_{j=1}^{\tau-1} P\Big[s_{t+j+1}, z_{t+j+1}, d_{t+j+1}, \mu_{t+j+1}\big|s_{t+j}, d_{t+j}, \mu_{t+j}, a_{t-\tau+j}\Big]\cdot P\Big[s_{t+1}, z_{t+1}, d_{t+1}, \mu_{t+1}\big|\mathscr{I}_t\Big]
\end{aligned}
$$

The result follows by noting

$$
P\Big[s_{t+1}, z_{t+1}, d_{t+1}, \mu_{t+1}\big|\mathscr{I}_t\Big] = \sum_{\mu_t} P\Big[s_{t+1}, z_{t+1}, d_{t+1}, \mu_{t+1}\big|s_t, d_t, \mu_t, a_{t-\tau}\Big]\cdot P[\mu_t|\mathscr{I}_t],
$$

and recalling the definition that $x_t = \{P[\mu_t|\mathscr{I}_t]\}$. □

We will make use of the following lemma in our proof of Propositions 19 and 20. Let

$$
\begin{aligned}
A_m(x) &\triangleq \tilde{h}_\tau \sum_{k=1}^{m} P\left[\sum_{j=1}^{\tau} d_j = \delta_k|x\right] - \tilde{p}_\tau \sum_{k=m+1}^{|\Delta_\tau|} P\left[\sum_{j=1}^{\tau} d_j = \delta_k|x\right] \\
B_m(x) &\triangleq \tilde{p}_\tau \sum_{k=m+1}^{|\Delta_\tau|} \delta_k P\left[\sum_{j=1}^{\tau} d_j = \delta_k|x\right] - \tilde{h}_\tau \sum_{k=1}^{m} \delta_k P\left[\sum_{j=1}^{\tau} d_j = \delta_k|x\right].
\end{aligned}
$$

Lemma 3. *$\mathbb{E}\left[g_\tau(y, d_1, \ldots, d_\tau)|x\right]$ is piecewise linear and convex in y, and has the form:*

$$
\mathbb{E}\left[g_\tau(y, d_1, \ldots, d_\tau)|x\right] = \begin{cases} A_0(x)y + B_0(x), & y \le \delta_1 \\ A_m(x)y + B_m(x), & \delta_m \le y \le \delta_{m+1} \\ A_{|\Delta_\tau|}(x)y + B_{|\Delta_\tau|}(x), & y \ge \delta_{|\Delta_\tau|}. \end{cases}
$$

Proof of Lemma 3. The inner function g_τ is convex and piecewise linear by canonical in-

ventory results. Convexity is preserved by expectation.

Piecewise linearity and the form above follow straightforwardly after noting:

$$\begin{aligned}\mathbb{E}\left[g_\tau(y,d_1,\ldots,d_\tau)|x\right] &= \sum_{d_1,\ldots,d_\tau} P\left[d_1,\ldots,d_\tau|x\right]\cdot\left[\tilde{h}_\tau\left(y-\sum_{j=1}^{\tau}d_j\right)^+ + \tilde{p}\left(\sum_{j=1}^{\tau}d_j - y\right)^+\right] \\ &= \sum_{\delta\in\Delta_\tau} P\left[\sum_{j=1}^{\tau}d_j=\delta|x\right]\cdot\left[\tilde{h}_\tau(y-\delta)^+ + \tilde{p}_\tau(\delta-y)^+\right].\end{aligned}$$

□

The proof of Proposition 19 follows along the lines of [31] Proposition 1.

Proof of Proposition 19. The proof is by induction on the value iterates in order to show that $v_n\big(u\vee y_\tau^*(x),x\big)$ is convex and non-decreasing, where $a\vee b\triangleq\max\{a,b\}$ and $v_n = H^{(d)}v_{n-1}$. Let $v_0=0$, so the induction hypothesis trivially holds. Suppose, for induction, that $v_n\big(u\vee y_\tau^*(x),x\big)$ is convex and non-decreasing in u, for all $x\in X$.

First case: $u\le y_\tau^*(x)$.

$$\begin{aligned}v_{n+1}(u,x) &= \min_{y\ge u}\left\{\beta^\tau\mathbb{E}\left[g_\tau(y,d_1,\ldots,d_\tau)|x\right]+\beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y-d',\lambda(d',z',x)\big)\right\} \\ &\le \beta^\tau\mathbb{E}\left[g_\tau(y_\tau^*(x),d_1,\ldots,d_\tau)|x\right]+\beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y_\tau^*(x)-d',\lambda(d',z',x)\big) \\ &= \beta^\tau\mathbb{E}\left[g_\tau\big(y_\tau^*(x),d_1,\ldots,d_\tau\big)|x\right]+\beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y_\tau^*\big(\lambda(d',z',x)\big),\lambda(d',z',x)\big)\end{aligned}$$

The last equality follows by applying the attainability condition. Note that the induction hypothesis implies that $v_n\big(\tilde{u},\lambda(d',z',x)\big) = v_n\big(y_\tau^*\big(\lambda(d',z',x),\lambda(d',z',x)\big)$ for all $\tilde{u}\le y_\tau^*\big(\lambda(d',z',x)$. We conclude that $v_{n+1}(u,x) = v_{n+1}\big(y_\tau^*(x),x\big)$ by the following ar-

gument:

$$
\begin{aligned}
v_{n+1}(u,x) &= \min_{y\geq u}\left\{\beta^{\tau}\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y-d',\lambda(d',z',x)\big)\right\}\\
&\geq \min_{y\geq u}\beta^{\tau}\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)\min_{y\geq u}v_n\big(y-d',\lambda(d',z',x)\big)\\
&= \beta^{\tau}\mathbb{E}\left[g_{\tau}\big(y_{\tau}^{*}(x),d_1,\ldots,d_{\tau}\big)|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y_{\tau}^{*}\big(\lambda(d',z',x)\big),\lambda(d',z',x)\big)\\
&= \beta^{\tau}\mathbb{E}\left[g_{\tau}(y_{\tau}^{*}(x),d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y_{\tau}^{*}(x)-d',\lambda(d',z',x)\big).
\end{aligned}
$$

Second case: $u > y_{\tau}^{*}(x)$. An upper bound on v_{n+1} follows straightforwardly by noting that choosing an order-up-to level equal to u is a feasible action. We construct a lower bound:

$$
\begin{aligned}
v_{n+1}(u,x) &= \min_{y\geq u}\left\{\beta^{\tau}\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(y-d',\lambda(d',z',x)\big)\right\}\\
&\geq \min_{y\geq u}\beta^{\tau}\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)\min_{y\geq u}v_n\big(y-d',\lambda(d',z',x)\big)\\
&= \beta^{\tau}\mathbb{E}\left[g_{\tau}(u,d_1,\ldots,d_{\tau})|x\right] + \beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(u-d',\lambda(d',z',x)\big).
\end{aligned}
$$

The last equality follows by convexity of $\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right]$ (Lemma 3) and the induction hypothesis on v_n. Thus, we conclude:

$$
\begin{aligned}
v_{n+1}(u,x) = \beta^{\tau}\mathbb{E}&\left[g_{\tau}(u\vee y_{\tau}^{*}(x),d_1,\ldots,d_{\tau})|x\right]\\
&+\beta\sum_{d',z'}\sigma(d',z'|x)v_n\big(u\vee y_{\tau}^{*}(x)-d',\lambda(d',z',x)\big).
\end{aligned}
$$

The result follows by taking the limit $n\to\infty$. □

Proof of Proposition 20. Suppose $y_{\tau}^{*}(x) = \delta_m$. By the convexity of the definition of $y_{\tau}^{*}(x)$ as the smallest minimizer of $\mathbb{E}\left[g_{\tau}(y,d_1,\ldots,d_{\tau})|x\right]$ and convexity, ordering up to δ_{m-1} and

δ_{m+1} yield the following set of inequalities:

$$A_{m-1}(x)\delta_{m-1}(x)\delta_{m-1} + B_{m-1}(x) > A_m(x)\delta_m + B_m(x)$$
$$A_{m+1}(x)\delta_{m+1} + B_{m+1}(x) \geq A_m(x)\delta_m + B_m(x).$$

The result follows by plugging in for the respective A and B functions and rearranging terms. □

Proof of Proposition 21. Note that

$$\sigma(d', z'|x) = \sum_{\tilde{z}} \xi(z'|\tilde{z})\tilde{\sigma}(d', \tilde{z}|x)$$
$$\lambda(\mu'|d', z', x) = \Lambda(d', z', \tilde{z}, x)\tilde{\lambda}(\mu'|d', \tilde{z}, x),$$

where

$$\sigma(d', z'|x) = \sum_{\mu', \mu} q(z'|\mu', \mu)P[d'|\mu', \mu]P[\mu'|\mu]x(\mu)$$
$$\tilde{\sigma}(d', \tilde{z}|x) = \sum_{\mu', \mu} \tilde{q}(\tilde{z}|\mu', \mu)P[d'|\mu', \mu]P[\mu'|\mu]x(\mu)$$
$$\Lambda(d', z', \tilde{z}, x) = \frac{\xi(z'|\tilde{z})\tilde{\sigma}(d', \tilde{z}|x)}{\sigma(d', z'|x)},$$

and $\{\Lambda(d', z', \tilde{z}, x)\}$ convex multipliers since $\sum_{\tilde{z}} \Lambda(d', z', \tilde{z}, x) = 1$.

Note that the concavity of v with respect to x ([48]) and Jensen's inequality imply

$$\sum_{\tilde{z}} \Lambda(d', z', \tilde{z}, x)\tilde{v}(u', \tilde{\lambda}(d', \tilde{z}, x)) \leq \tilde{v}(u', \lambda(d', \tilde{z}, x)). \tag{B.1}$$

Proof is by induction on the value iteration iterates. The base case $\tilde{v}_0 \leq v_0$ is satisfied by assumption. Suppose, for induction, that $\tilde{v}_n \leq v_n$, and that $v_{n+1} = Hv_n$ and $\tilde{v}_{n+1} = \tilde{H}\tilde{v}_n$. For

compactness of notation, let $d_{[1:\tau]} \triangleq (d_1, \ldots, d_\tau)$.

$$
\begin{aligned}
v_{n+1}(u,x) &= \min_{y \geq u} \left\{ \mathbb{E}\big[g_\tau(y, d_{[1:\tau]})|x\big] + \beta \sum_{d',z'} \sigma(d', z'|x) v_n\big(y - d', \lambda(d', z', x)\big) \right\} \\
&\geq \min_{y \geq u} \left\{ \mathbb{E}\big[g_\tau(y, d_{[1:\tau]})|x\big] + \beta \sum_{d',z'} \sigma(d', z'|x) \tilde{v}_n\big(y - d', \lambda(d', z', x)\big) \right\} \\
&\quad \text{by the induction hypothesis} \\
&\geq \min_{y \geq u} \left\{ \mathbb{E}\big[g_\tau(y, d_{[1:\tau]})|x\big] + \beta \sum_{d',z',\tilde{z}} \sigma(d', z'|x) \Lambda(d', z', \tilde{z}, x) \tilde{v}_n\big(y - d', \tilde{\lambda}(d', \tilde{z}, x)\big) \right\} \\
&\quad \text{by Equation B.1} \\
&= \min_{y \geq u} \left\{ \mathbb{E}\big[g_\tau(y, d_{[1:\tau]})|x\big] + \beta \sum_{d',z',\tilde{z}} \xi(z'|\tilde{z}) \tilde{\sigma}(d', \tilde{z}|x) \tilde{v}_n\big(y - d', \tilde{\lambda}(d', \tilde{z}, x)\big) \right\} \\
&\quad \text{by plugging in for } \Lambda \\
&= \min_{y \geq u} \left\{ \mathbb{E}\big[g_\tau(y, d_{[1:\tau]})|x\big] + \beta \sum_{d',\tilde{z}} \tilde{\sigma}(d', \tilde{z}|x) \tilde{v}_n\big(y - d', \tilde{\lambda}(d', \tilde{z}, x)\big) \right\} \\
&= \tilde{v}_{n+1}(u,x)
\end{aligned}
$$

The result follows by taking the limit as $n \to \infty$. □

Proof of Proposition 22. We've shown that for the τ'-lagged problem, it is sufficient for the DM to make decisions at epoch t on the basis of (u_t, x_t). Suppose that the DM is a clairvoyant for $\tau' - \tau$ epochs into the future, so that the DM makes decisions at epoch t on the basis of $(u_t, x_t, d_{t+1}, \ldots, d_{t+\tau'-\tau}, z_{t+1}, \ldots, z_{t+\tau'-\tau})$, *i.e.* the DM knows the realizations of the demand and AOD observations for $\tau' - \tau$ epochs into the future.

Consider the single period cost function for this new clairvoyant DM for the τ'-lagged

problem. Let $d_{[1:\tau'-\tau]} = (d_1, \ldots, d_{\tau'-\tau})$ and $z_{[1:\tau'-\tau]}$ be defined likewise.

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{h}_{\tau'}\left(y - \sum_{j=1}^{\tau'} d_j\right)^+ + \tilde{p}_{\tau'}\left(\sum_{j=1}^{\tau'} d_j - y\right)^+ | d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x\right] \\
&= \mathbb{E}\left[\tilde{h}_{\tau'}\left(\bar{y} - \sum_{j=\tau'-\tau+1}^{\tau'} d_j\right)^+ + \tilde{p}_{\tau'}\left(\sum_{j=\tau'-\tau+1}^{\tau'} d_j - \bar{y}\right)^+ | d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x\right] \\
&\quad \text{where } \bar{y} = y - \sum_{j=1}^{\tau'-\tau} d_j \geq u - \sum_{j=1}^{\tau'-\tau} d_j \\
&= \mathbb{E}\left[\tilde{h}_{\tau'}\left(\bar{y} - \sum_{j=\tau'-\tau+1}^{\tau'} d_j\right)^+ + \tilde{p}_{\tau'}\left(\sum_{j=\tau'-\tau+1}^{\tau'} d_j - \bar{y}\right)^+ | \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right] \\
&= \mathbb{E}\left[\tilde{h}_{\tau'}\left(\bar{y} - \sum_{j=\tau'-\tau+1}^{\tau'} d_j\right)^+ + \tilde{p}_{\tau'}\left(\sum_{j=\tau'-\tau+1}^{\tau'} d_j - \bar{y}\right)^+ | \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right]
\end{aligned}
$$

The first equality follows by noting that the clairvoyant demand information allows us to reformulate the cost function around the remaining random variables $(d_{\tau'-\tau}, \ldots, d_{\tau'})$ and a new stocking decision $\bar{y}$. The second equality follows by noting than an equivalent sufficient statistic for the τ'-lagged clairvoyant is $\left(u - \sum_{j=1}^{\tau'-\tau} d_j, \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right)$, where $\lambda^{\tau'-\tau}$ is the $(\tau' - \tau)$-fold Bayesian update on the basis of the clairvoyant demand realizations $d_{[1:\tau'-\tau]}$ and AOD observations $z_{[1:\tau'-\tau]}$.

Let $v^*_{\tau',c}$ to be the fixed point of $H_c^{\tau'}$ the Bellman operator of the τ'-lagged clairvoyant problem (defined below), and

$$
\begin{aligned}
&H_c^{\tau'} v\left(u, x, d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}\right) = \\
&\quad \min_{\bar{y} \geq u - \sum_{j=1}^{\tau'-\tau} d_j} \left\{ \mathbb{E}\left[\tilde{h}_{\tau'}\left(\bar{y} - \sum_{j=\tau'-\tau+1}^{\tau'} d_j\right)^+ + \tilde{p}_{\tau'}\left(\sum_{j=\tau'-\tau+1}^{\tau'} d_j - \bar{y}\right)^+ | \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right] \right. \\
&\quad + \beta \sum_{\substack{d_{\tau'-\tau+1}, \\ z_{\tau'-\tau+1}}} \sigma\left(d_{\tau'-\tau+1}, z_{\tau'-\tau+1} | \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right) \times \cdots \\
&\quad \left. \times v\left(\bar{y} - d_1, \lambda(d_1, z_1, x), d_{[2:\tau'-\tau+1]}, z_{[2:\tau'-\tau+1]}\right) \right\}.
\end{aligned}
$$

Note that this is the τ-lagged Bellman operator, except with the holding cost and penalty

cost per unit dependent on τ' instead of τ. By using the relationships $\beta^{\tau'-\tau}\tilde{h}_\tau \leq \tilde{h}_{\tau'}$ and $\beta^{\tau'-\tau}\tilde{p}_\tau \leq \tilde{p}_{\tau'}$, and correcting for the discount factor, we have

$$v_\tau^*\left(u - \sum_{j=1}^{\tau'-\tau} d_j, \lambda^{\tau'-\tau}(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x)\right) \leq \beta^{\tau-\tau'} v_{\tau',c}^*\left(u, x, d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}\right).$$

However, since the clairvoyance assumption is an information relaxation, by weak duality ([7]) we have the following:

$$v_\tau^*\left(u - \sum_{j=1}^{\tau'-\tau} d_j, \lambda^{\tau'-\tau}\big(d_{[1:\tau'-\tau]}, z_{[1:\tau'-\tau]}, x\big)\right) \leq \beta^{\tau-\tau'} v_{\tau'}^*(u, x).$$

□

B.3 Alternative Formulations

There are some atypical characteristics to the optimality equation in (3.3) — namely, the dependence of $C(s, d, a, a_{-\tau})$ on the τ-lagged action $a_{-\tau}$. We seek to reformulate the problem in a more familiar way by shifting the perspective of the DM at each decision epoch.

We do so by the following line of reasoning:

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t C(s_t,d_t,a_t,a_{t-\tau})|\mathscr{I}_0\right] = \mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t\left[c_\tau a_t + h_\tau(s_t + a_{t-\tau} - d_t)^+ + p_\tau(d_t - s_t - a_{t-\tau})^+\right]|\mathscr{I}_0\right] \tag{B.2}$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t c_\tau a_t + \sum_{t=-\tau}^{\infty}\beta^{t+\tau}\left[h_\tau(s_{t+\tau} + a_t - d_{t+\tau})^+ + p_\tau(d_{t+\tau} - s_{t+\tau} - a_t)^+\right]|\mathscr{I}_0\right] \tag{B.3}$$

$$= \mathbb{E}\left[\sum_{t=-\tau}^{-1}\beta^{t+\tau}\left[h_\tau(s_{t+\tau} + a_t - d_{t+\tau})^+ + p_\tau(d_{t+\tau} - s_{t+\tau} - a_t)^+\right]|\mathscr{I}_0\right] \tag{B.4}$$

$$+ \mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t c_\tau a_t + \sum_{t=0}^{\infty}\beta^{t+\tau}\left[h_\tau(s_{t+\tau} + a_t - d_{t+\tau})^+ + p_\tau(d_{t+\tau} - s_{t+\tau} - a_t)^+\right]|\mathscr{I}_0\right] \tag{B.5}$$

$$= D + \mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t c_\tau a_t + \beta^{t+\tau}\left[h_\tau(s_{t+\tau} + a_t - d_{t+\tau})^+ + p_\tau(d_{t+\tau} - s_{t+\tau} - a_t)^+\right]|\mathscr{I}_0\right], \tag{B.6}$$

where D is the term in line (B.4).

Minimizing the expectation in (B.6) is equivalent to minimizing the total discounted costs, due to the independence of the cost in the first τ epochs with respect to control. This corresponds to the Bellman optimality equation $v = H^{(b)}v$, where $H^{(b)}$ is defined as follows, and we let $a_{[-1:-\tau]} = (a_{-1}, \ldots, a_{-\tau})$:

$$H^{(b)}v(s,d,a_{[-1:-\tau]},x) = \min_a \Bigg\{ c_\tau a + \beta^\tau \mathbb{E}\left[h_\tau(s_\tau + a - d_\tau)^+ + p_\tau(d_\tau - s_\tau - a)^+|s,d,a_{[-1:-\tau]},x\right] + \beta \sum_{d',z'} \sigma(d',z'|x) v\left(s + a_{-\tau} - d, d', a, a_{-1}, \ldots, a_{-\tau+1}, \lambda(d',z',x)\right)\Bigg\}. \tag{B.7}$$

At first, this reformulation may not appear to give much computational advantage using traditional solution procedures such as value iteration or policy iteration, since computing the expectation in the single period cost may itself be computationally expensive. However, we will show that there are *analytical* advantages to the reformulation that can be used to

develop specialized solution procedures.

Note that the inner function $g_\tau(s_\tau, a, d_\tau) = h(s_\tau + a - d_\tau)^+ + p(d_\tau - s_\tau - a)^+$ is familiar from canonical inventory problems. The function g_τ is piecewise linear and convex, and these structural properties are preserved under expectation. Further, g_τ is dependent on the action of the current decision epoch, rather than on the past action, which elucidates the dependence on action in the optimization within Equation B.7.

There may be interpretative advantages to different, equivalent formulations. In the original formulation, costs are accounted for *on delivery*. In this new reformulation, costs are accounted for at the time of decision by projecting forward τ decision epochs.

We may further modify the formulation by defining $y_t = s_t + \sum_{j=1}^{\tau} a_{t-j} + a_t - d_t$, the total amount of inventory possessed through the interval $[t, t+\tau]$, and noting that $s_{t+\tau} = y_t - a_t - \sum_{j=0}^{\tau-1} d_{t+j}$. If we let $u_t = y_t - a_t$ be the inventory position through interval $[t, t+\tau]$ before ordering, then we have that $u_{t+1} = y_t - d_{t+1}$, which is familiar as the inventory difference equation under backlogging. The resulting optimality equation is $v = H^{(c)}v$, where $H^{(c)}$ is defined to be:

$$\begin{aligned} H^{(c)}v(u,x) = \min_{y \geq u} \Bigg\{ & c_\tau(y-u) + \beta^\tau \mathbb{E}\left[h_\tau\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + p_\tau\left(\sum_{j=1}^{\tau} d_j - y\right)^+ | x \right] \\ & + \beta \sum_{d', z'} \sigma(d', z'|x) v\big(y - d', \lambda(d', z', x)\big) \Bigg\}. \end{aligned} \tag{B.8}$$

Then we note that the expectation of the single period cost with respect to $d_1, \ldots, d_\tau$ only depends on x. Thus, we have projected out the dependence on the inventory stock level in the single period cost function. This step is a transformation analogous to the transformations used for determining (1) base stock optimal policies when $\tau = 1$ in [31], and (2) the optimality of myopic policies in [47] for MDPs and in [5] for M-POMDPs.

Finally, denote let the operator $\tilde{H}$ from Equation 3.4 be denoted $H^{(d)}$, as well, to denote the sequence of the reformulation.

B.4 Relationships Between the Fixed Points

In this appendix, we detail the relationship between the fixed points of the operators $\{H^{(i)} : i = a, b, c, d\}$, which we denote by $v^{(i)}$ the unique fixed point of $H^{(i)}$, that define alternative dynamic programming formulations of the inventory control problem of Section 3.2.

For ease of reference, we provide the different operator definitions below, where we suppress t from the notation for compactness:

$$
\begin{aligned}
H^{(a)}v^{(a)}(s, d, a_{[-1:-\tau]}, x) &= \min_{a} \Bigg\{ c_\tau a_{-\tau} + h_\tau(s + a_{-\tau} - d)^+ + p_\tau(d - s - a_{-\tau})^+ + \\
&\qquad \beta \sum_{d', z'} \sigma(d', z'|x) v^{(a)}\big(s - d + a_{-\tau}, d', a, \ldots, a_{-\tau+1}, \lambda(d', z', x)\big) \Bigg\} \\
H^{(b)}v^{(b)}(s, d, a_{[-1:-\tau]}, x) &= \min_{a} \Bigg\{ c_\tau a + \beta^\tau \mathbb{E}\big[h_\tau(s_\tau + a - d_\tau)^+ + p_\tau(d_\tau - s_\tau - a)^+ | s, d, a_{[-1:-\tau]}, x\big] \\
&\qquad + \beta \sum_{d', z'} \sigma(d', z'|x) v^{(b)}\big(s + a_{-\tau} - d, d', a, a_{-1}, \ldots, a_{-\tau+1}, \lambda(d', z', x)\big) \Bigg\} \\
H^{(c)}v^{(c)}(u, x) &= \min_{y \geq u} \Bigg\{ c_\tau(y - u) + \beta^\tau \mathbb{E}\left[h_\tau\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + p_\tau\left(\sum_{j=1}^{\tau} d_j - y\right)^+ | x \right] \\
&\qquad + \beta \sum_{d', z'} \sigma(d', z'|x) v^{(c)}\big(y - d', \lambda(d', z', x)\big) \Bigg\} \\
H^{(d)}v^{(d)}(u, x) &= \min_{y \geq u} \Bigg\{ \mathbb{E}\left[\tilde{h}_\tau\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + \tilde{p}_\tau\left(\sum_{j=1}^{\tau} d_j - y\right)^+ | x \right] \\
&\qquad + \beta \sum_{d', z'} \sigma(d', z'|x) v^{(d)}\big(y - d', \lambda(d', z', x)\big) \Bigg\}.
\end{aligned}
$$

The relationship between the (a) and (b) formulations is straightforward, as detailed in Section 3.3.1: $v^{(a)} = D + v^{(b)}$, where

$$
D = \mathbb{E}\left[\sum_{t=-\tau}^{-1} \beta^{t+\tau} \big[h_\tau(s_{t+\tau} + a_t - d_{t+\tau})^+ + p_\tau(d_{t+\tau} - s_{t+\tau} - a_t)^+ \big] | \mathscr{I}_0 \right].
$$

The (b) and (c) formulations are equivalent when $u = s - d + \sum_{j=1}^{\tau} a_{-j}$. Thus,

$$v^{(b)}(s, d, a_{[-1:-\tau]}, x) = v^{(c)}\left(s - d + \sum_{j=1}^{\tau} a_{-j}, x\right).$$

The relationship between the (c) and (d) formulations requires some work. Note that since $a = (a)^+ - (-a)^+$ for any $a \in \mathbb{R}$,

$$y = \mathbb{E}\left[\sum_{j=1}^{\tau} d_j + \left(y - \sum_{j=1}^{\tau} d_j\right)^+ - \left(-\left(y - \sum_{j=1}^{\tau} d_j\right)\right)^+ |x\right]. \tag{B.9}$$

Plugging into the equation $v^{(c)} = H^{(c)}v^{(c)}$:

$$\begin{aligned} H^{(c)}v^{(c)}(u, x) = &-c_\tau u + c_\tau \mathbb{E}\left[\sum_{j=1}^{\tau} d_j|x\right] \\ &+ \min_{y \geq u}\left\{\mathbb{E}\left[(c_\tau + \beta^\tau h_\tau)\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + (c_\tau + \beta^\tau p_\tau)\left(\sum_{j=1}^{\tau} d_j - y\right)^+ |x\right]\right. \\ &\left. + \beta \sum_{d', z'} \sigma(d', z'|x) v^{(c)}\big(y - d', \lambda(d', z', x)\big)\right\}. \end{aligned}$$

Let $\tilde{v}(u, x) = v^{(c)}(u, x) + cu$ and plug into the fixed point equation $v^{(c)} = H^{(c)}v^{(c)}$:

$$\begin{aligned} \tilde{v}(u, x) = c_\tau \mathbb{E}\left[\sum_{j=1}^{\tau} d_j|x\right] + \min_{y \geq u}\left\{\mathbb{E}\left[(c_\tau + \beta^\tau h_\tau)\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + (c_\tau + \beta^\tau p_\tau)\left(\sum_{j=1}^{\tau} d_j - y\right)^+ |x\right]\right. \\ \left. + \beta \sum_{d', z'} \sigma(d', z'|x)\tilde{v}\big(y - d', \lambda(d', z', x)\big) - \beta c_\tau (y - d')\right\}. \end{aligned}$$

Finally, if we substitute once more for y from Equation B.9, we get:

$$\begin{aligned} \tilde{v}(u, x) = &\, c_\tau(1 - \beta)\mathbb{E}\left[\sum_{j=1}^{\tau} d_j|x\right] + \beta c_\tau \mathbb{E}[d'|x] \\ &+ \min_{y \geq u}\left\{\mathbb{E}\left[(c_\tau + \beta^\tau h_\tau)\left(y - \sum_{j=1}^{\tau} d_j\right)^+ + (c_\tau + \beta^\tau p_\tau)\left(\sum_{j=1}^{\tau} d_j - y\right)^+ |x\right]\right. \\ &\left. + \beta \sum_{d', z'} \sigma(d', z'|x)\tilde{v}\big(y - d', \lambda(d', z', x)\big)\right\}. \end{aligned}$$

Thus, we get that $\tilde{v} = K + H^{(d)}\tilde{v}$, where $K = c_\tau(1-\beta)\mathbb{E}\left[\sum_{j=1}^{\tau} d_j|x\right] + \beta c_\tau \mathbb{E}[d'|x]$. It is then straightforward to show that $v^{(d)} = \tilde{v} - \frac{K}{1-\beta}$.

Altogether, the relationship between $v^{(a)}$ and $v^{(d)}$ is as follows:

$$v^{(a)}(s, d, a_{-1}, \ldots, a_{-\tau}, x) = D + v^{(d)}(u, x) - cu + \frac{K}{1-\beta},$$

where $u = s - d + \sum_{j=1}^{\tau} a_{-j}$ and D, K are defined above.

B.5 On Stock-outs

B.5.1 Stock-out Robust Policies

In the rest of this paper, we have been concerned with policies that minimize expected total discounted costs. However, the DM may have additional objectives that reflect the DM's risk profile, *e.g.* hedging against the upper tail of the total cost distribution or maintaining a particular service level at all times. In this subsection, we develop policies that consider the frequency with which orders are backlogged (and correspondingly, robust to dips in service level). The true stock-out penalty term p_τ should capture the per-unit cost of backlogging, incorporating all risk tolerances to stock-outs. However, it is often difficult in practice to determine a "true value" of p_τ that incorporates all of these notions of risk, because it may be unnatural for the DM to "price" their risk aversion to stock-outs on a per-unit basis. It may be more natural to express risk aversion in terms of "maintaining a particular service level" or "keeping the probability of stock-outs below a certain threshold". Therefore, we consider p_τ to reflect all other per-unit costs pertaining backlogging and treat the DM's risk aversion in terms of a chance constraint. We seek to illuminate the relationship between this chance constrained formulation and our original formulation by viewing the term p_τ as a way to control the stock-out level.

We consider base stock levels, $y_{\theta_p,\tau}(x)$, parametrized by hypothetical stock-out penalty

values θ_p, that are the smallest (and hence unique) minimizer such that:

$$y_{\theta_p,\tau}(x) \in \arg\min_y \left\{ \mathbb{E}\left[\tilde{h}_\tau \left(y - \sum_{j=1}^{\tau} d_j \right)^+ + \theta_p \left(\sum_{j=1}^{\tau} d_j - y \right)^+ | x \right] \right\}.$$

Let $\pi_{\theta_p,\tau}(u,x) \triangleq \max\{y_{\theta_p,\tau}(x) - u, 0\}$ for all u, x, the base stock policy generated by θ_p. The next proposition states that, given the same underlying uncertainty realization, the empirical stock-out rate (for any finite horizon) will be smaller under a base stock policy generated by a higher penalty term, θ_p.

Proposition 25. *Suppose $\theta_p < \theta'_p$, let T be a finite horizon, and let ω be a realization of the underlying uncertainty. Then,*

$$\frac{1}{T}\sum_{t=0}^{T} \mathbf{1}\{s_t(\omega) < 0\} \geq \frac{1}{T}\sum_{t=0}^{T} \mathbf{1}\{s'_t(\omega) < 0\},$$

where $\{s_t\}$ are generated by π_p and $\{s'_t\}$ are generated by $\pi_{\theta'_p}$.

Proof of Proposition 25. Suppose, without loss of generality, that $s_0 = s'_0$. The dynamics for $\{s_t\}$ and $\{s'_t\}$ are described functionally in terms of the like realizations of demand $\{d_t(\omega)\}$ (the ω is included to denote that these are realizations of demand rather than random variables). Further, the belief evolution $\{x_t\}$ and $\{x'_t\}$ are due to common realizations of demand $\{d_t(\omega)\}$ and AOD observations $\{z_t(\omega)\}$. Thus $x_t = x'_t$ for all t. The result follows by simply noting that $y_p(x) \leq y_{p'}(x)$ for all $x \in X$. Thus, if $s'_{t+1}(\omega) = y_{p'}(x_t) - d_t(\omega) < 0$, then $s_{t+1}(\omega) < 0$, but the converse is not necessarily true. □

The finite horizon in Proposition 25 resembles our numerical analysis in Section 3.5, in which we will approximate the infinite horizon value of policies using Monte Carlo simulation over a sufficiently long finite horizon. Proposition 25 formalizes the intuition that, given the same sample path, the realized stock-out rate will be less under a base stock policy that penalizes stock-outs more will always generate a stock-out rate at most as high as a base stock policy that penalizes stock-outs less.

Increasing the parameter θ_p used to generate policy $\pi_{\theta_p,\tau}$ decreases the stock-out rate path-wise, by Proposition 25, but it also generates policies are suboptimal due to the mismatch between the true stock-out penalty and the penalty used to generate the policy $\pi_{\theta_p,\tau}$. We refer to the marginal difference in value, given some initialization (u_0, x_0), to be $v^*_\tau(u_0, x_0) - v^{\pi_{\theta_p,\tau}}(u_0, x_0)$ as the *marginal value of stock-out robustness*, which we note will typically be ≤ 0.

Relationship to Chance Constraints. We seek to illuminate the intuition behind generating these stock-out robust policies by relating these policies to a chance constrained version of the problem. Consider the following formulation of the problem, in which the DM wants the *a priori* probability of a stock-out to be less than or equal to α (the service level greater than $1 - \alpha$) in each decision epoch.

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}\left[\sum_{t=0}^{T} \beta^t \big(\tilde{h}_\tau(s_t + a_{t-\tau} - d_t)^+ + \tilde{p}_\tau(d_t - s_t - a_{t-\tau})^+\big) | \mathscr{I}_0\right] \\ \text{s.t.} \quad & P\left[s_t < 0 | \mathscr{I}_0\right] \leq \alpha, \quad \forall t \end{aligned}$$

The constraint $P\left[s_t < 0 | \mathscr{I}_0\right] \leq \alpha$ for all t is a chance constraint, and under this formulation of the problem it is a hard constraint (the service level must be above $1 - \alpha$). Note that we can rewrite the constraint as $\mathbb{E}[\mathbf{1}\{s_t < 0\} | \mathscr{I}_0] \leq \alpha$, and then consider the following Lagrangian relaxation with penalties $\lambda = \{\lambda_t\} \geq 0$ (an abuse of notation, not to be confused with the posterior distribution λ).

$$\begin{aligned} & \min_{\pi} \left\{ \mathbb{E}\left[\sum_{t=0}^{T} \beta^t \big(\tilde{h}_\tau(s_t + a_{t-\tau} - d_t)^+ + \tilde{p}_\tau(d_t - s_t - a_{t-\tau})^+\big) | \mathscr{I}_0\right] - \sum_{t=0}^{T} \lambda_t \big(\alpha - \mathbb{E}\left[\mathbf{1}\{s_t < 0\} | \mathscr{I}_0\right]\big) \right\} \\ & = \min_{\pi} \left\{ \mathbb{E}\left[\sum_{t=0}^{T} \beta^t \left(\tilde{h}_\tau(s_t + a_{t-\tau} - d_t)^+ + \tilde{p}_\tau(d_t - s_t - a_{t-\tau})^+ + \frac{\lambda_t}{\beta^t} \mathbf{1}\{s_{t+1} < 0\}\right) | \mathscr{I}_0\right] + \text{cons.} \right\} \\ & = \min_{\pi} \left\{ \mathbb{E}\left[\sum_{t=0}^{T} \beta^t \left(\tilde{h}_\tau(s_t + a_{t-\tau} - d_t)^+ + \left(\tilde{p}_\tau + \frac{\lambda_t}{\beta^t (d_t - s_t - a_{t-\tau})}\right)(d_t - s_t - a_{t-\tau})^+\right) | \mathscr{I}_0\right] + \text{cons.} \right\} \\ & \approx \min_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \beta^t \left(\tilde{h}_\tau(s_t + a_{t-\tau} - d_t)^+ + \bar{p}_\tau(d_t - s_t - a_{t-\tau})^+\right) | \mathscr{I}_0\right], \quad \text{where } \bar{p}_\tau \geq \tilde{p}_\tau. \end{aligned}$$

Hence, increasing the penalty term acts like including a Lagrangian multiplier in the Lagrangian relaxation of the chance constrained formulation of the problem. Thus, we may approximate the optimal policies generated in the chance constrained problem by increasing the penalty term and solving the original problem. Since base stock policies are optimal in the unconstrained original formulation of the problem (under an attainability condition), we generate base stock policies by varying the underage penalty as approximations to the chance-constrained problem.

B.5.2 Numerical Analysis

Here we present the numerical results pertaining to θ_p and stock-outs for the numerical example of Section 3.5. We denote by SO_θ, the number of observed stock-outs in the Monte Carlo simulation due to policy π_θ:

$$\mathrm{SO}_\theta = \sum_{n=1}^{N_{sim}} \sum_{t=0}^{T} \mathbf{1}\{s_t^n < 0\}.$$

Effect of θ_p on discounted costs. Recall that we discussed the relationship between expected total discounted costs (the mean value, v_θ) and the per-unit underage penalty parameter θ_p as the *value of stock-out robustness*. We showed in Proposition 19 that base stock policies are optimal under an attainability condition that we found to be quite robust in this numerical example. Thus, we expect the SVM-generated base stock policy with parameter θ_p equal to the true per-unit underage penalty cost $p = 3$ to have the lowest mean value, even though our policy generation method is approximate in nature and subject to classification error (due to the SVM partitioning method) and simulation error (due to the Monte Carlo policy evaluation). All other per-unit underage penalty parameters, $\theta_p = 4, 5, 6$, generate suboptimal base stock policies due to the uniqueness of the optimal base stock levels, and thus we expect higher mean values. Additionally, we discussed in Section 3.4.5 this trade-off between sub-optimality of policies generated by θ_p larger than $p = 3$

under the term the *value of stock-out robustness*.

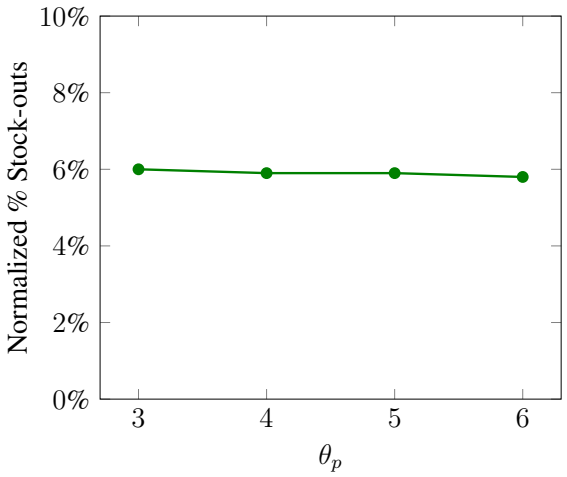
Following Equation 3.9, we estimate the average effects of varying θ_p and θ_q while holding τ and θ_M constant, by normalizing to $\theta_p = 3$ (the true penalty term) and $\theta_q = 0$ (perfect AOD information) and computing $e^{(b_{\theta_p}+b_{\theta_q})}$. In Figure B.2 we can see that increasing θ_p yields a significant and greater degree of sub-optimality in terms of long-run costs for our example. Choosing the policy parametrized $\theta_p = 3$, aligned with the true per-unit underage cost p, is estimated to be worth 5.4% less in long run costs than $\theta_p = 4$, 9.4% less in long run costs than $\theta_p = 5$, and 12.7% less in long run costs for $\theta_p = 6$. Further, in Figure B.2(c) we again see that for these parameter values the mean value is concave, which means that there is increasing marginal effect on decreasing long run costs as θ_p approaches $p = 3$.

Effect of θ_p ***on variance.*** In this example, we see less drastic effects on standard error in changing θ_p than in changing AOD information quality. These results are depicted in Figures B.1c, B.1d, and B.1e. Increasing the stock-out penalty term has modest effects in reducing variance of long run costs. However, for this example the effects are concentrated to increasing θ_p over the true penalty, $p = 3$. This result may be due to the nature of this particular example, in which increasing the penalty from $\theta_p = 3$ to $\theta_p = 4$ leads to a more significant increase in base stock levels in general than increasing $\theta_p = 4$ to $\theta_p = 5$. These effects also might alternatively be due to the narrow range of θ_p that were tested.

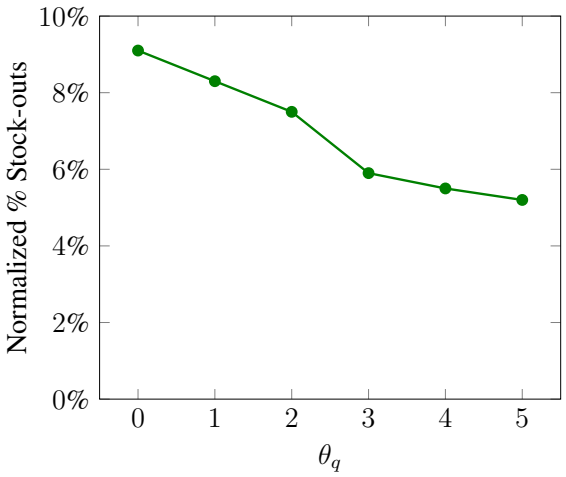
On Stock-outs. For the DM concerned with service level, the third metric that we measure is stock-out rates.

Somewhat unexpectedly the step-wise log-linear regression with categorical regressors (Equation 3.8) yielded some θ_p values with borderline statistically significant, so we ran the same regression replacing the categorical θ_p regressors with the numeric $\log(\theta_p)$ regressor, the results of which are in Table B.1. This log-log regression amounts to an implicit assumption of constant elasticity of stock-outs with respect to stock-out penalty parameter θ_p.

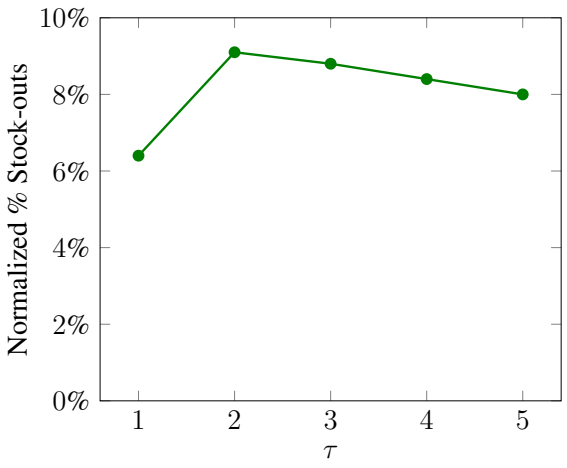
The predicted effects on stock-outs are depicted in Figure B.6, and each graph is nor-



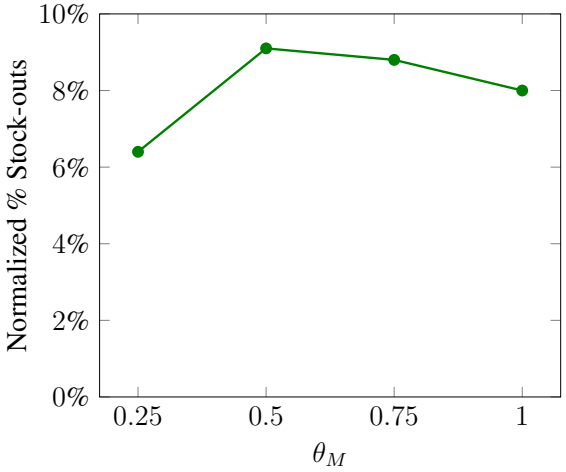
(a) Predicted percentage of stock-outs for various underage penalty parameters, θ_p, normalized to the median stock-out rate at $\theta_p = 1$.



(b) Predicted percentage of stock-outs for various AOD parameters, θ_q, normalized to the median stock-out rate at $\theta_q = 0$.



(c) Predicted percentage of stock-outs for various lead times, τ, normalized to the median stock-out rate at $\tau = 1$.



(d) Predicted percentage of stock-outs for various modulation parameters, θ_M, normalized to the median stock-out rate at $\theta_M = 0.25$.

Figure B.6: Marginal effects on stock-outs.

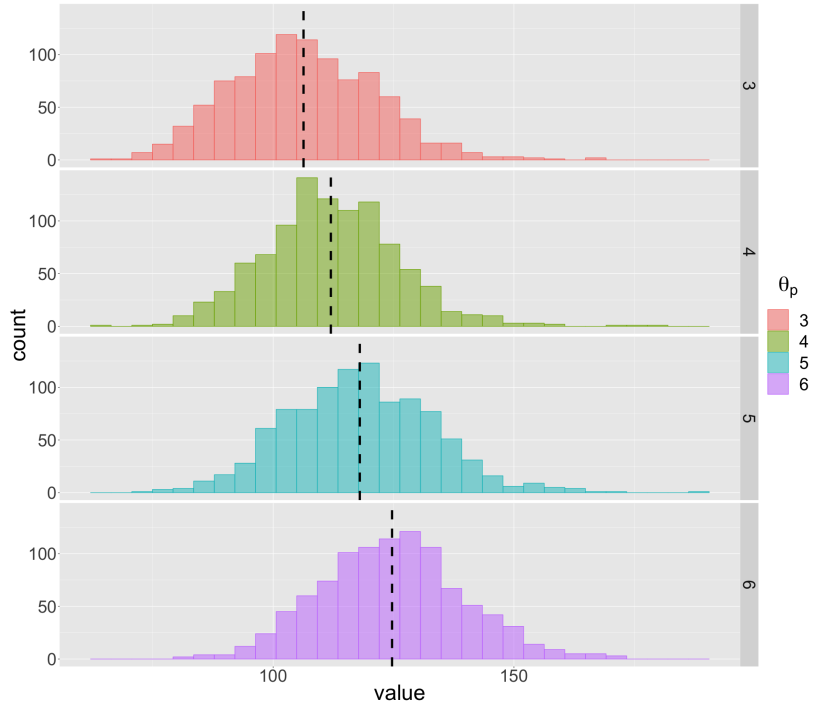


Figure B.7: Histogram of simulated values by θ_p, for fixed $\theta_M = 0.5$, $\theta_q = 2$, $\tau = 2$. The dashed lines represent the sample means.

malized to median stock-out rates. Our first observation is that the system is relatively stable with respect to stock-outs. The majority of the sampled stock-out rates from our Monte Carlo simulation were less than 10%, and the highest sampled stock-out rate was 11.2%.

As Figure B.6a shows, increasing θ_p has an effect of decreasing stock-outs as we discussed in Section 3.4.5. However, for this numerical example and over the range of θ_p considered, this effect is relatively insignificant. Since the effect of increasing θ_p yielded notable increases in discounted costs, as we discussed earlier, deviating from $\theta_p = 3$ is inadvisable in this particular example. We anticipate in other examples, the effect of increasing θ_p on reducing stock-outs might be more pronounced.

Counter-intuitively, we see in Figure B.6b, that better AOD information yielded an increase in stock-outs for this numerical example. There are different potential explanations for why we observe this phenomenon. One potential explanation is that the presence of uncertainty due to poor information quality may cause the order-up-to levels to increase,

as the DM hedges against uncertainty by ordering more inventory — leading to a lower stock-out rate. Essentially, in this interpretation, better AOD information allows the DM to practice *leaner* inventory management, with lower inventory levels. This explanation seems especially plausible when the DM is in a low-demand state of the economy, but believes (due to information uncertainty) they might be in a high-demand state of the economy. Counter to this explanation, if the DM is in a high-demand state state of the economy, but believes they might be in a low-demand state, then the DM might order less. The aggregate effect on stock-outs would thus depend on the balance between these two effects.

Similarly to what we have observed with our discounted costs and standard error, as θ_M either approaches 0 or 1, stock-outs decrease. This reinforces the observation that more stable and predictable macroeconomic environments yield better system performance, in terms each of our metrics.

Finally, the effect of τ on stock-outs is non-monotone in our example. Thus, we cannot make a plausible generalization that increasing or decreasing lead times will improve or exacerbate stock-outs. Altogether, the effect of managerial decisions on stock-outs should be evaluated on a case-by-case basis since simple rules-of-thumb pertaining to the relationship between stock-outs and some of our input parameters remain elusive.

APPENDIX C

GENERATING TRUST IN DEVELOPMENT PROCESSES USING ROBUST, DATA-DRIVEN MARKOV GAMES: AN APPLICATION TO PRESTIGE

C.1 Determining the Transition Probabilities

In determining the transition probabilities in this section, we suppose that the actions are feasible, namely that the context for the actions are satisfied. We first determine the transition probabilities for *game nodes*. Note

$$\begin{aligned} P\left[S_{t+1}|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] &= \sum_{S_t' \in \{0,1\}^N} P\left[S_{t+1}, S_t'|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] \\ &= \sum_{S_t' \in \{0,1\}^N} P\left[S_{t+1}|S_t', A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] P\left[S_t'|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] \end{aligned}$$

The terms on the right hand side decompose into the following

$$\begin{aligned} P\left[S_t'|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] &= \prod_{n \in N} P\left[s_t'(n)|S_t, A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] \\ P\left[S_{t+1}|S_t', A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right] &= \prod_{n \in N} P\left[s_{t+1}(n)|S_t', A_t^{\mathcal{D}}, A_t^{\mathcal{A}}\right]. \end{aligned}$$

We can compute these constituent probabilities for $s_t'(n)$ using the dynamics rules in Section 4.3. Suppose agent k can take action $a_t^k(n)$ at epoch t and node n that has three components: $\tau_t^k(n)$ corresponding to a Take move, $\varphi_t^k(n)$ corresponding to a Protect move, and $\omega_t^k(n)$ corresponding to an Observe move. Take moves take values in $\{0, 1\}$, Protect moves take values in $[0, 1]$, and Observe moves take values in $\{0, 1\}$. Take and Protect moves are

the only actions that affect the transition dynamics here.

$$
\begin{aligned}
P\left[s'_t(n) = s_t(n) | \tau_t^{\mathcal{D}}(n) = 0\right] &= 1 \\
P\left[s'_t(n) = \mathcal{D} | s_t(n) = \mathcal{A}, \tau_t^{\mathcal{D}}(n) = 1, \varphi_t^{\mathcal{A}}(n)\right] &= P^{\mathcal{D}}\left(\tau_t^{\mathcal{D}}(n)\right) \cdot \left(1 - \varphi_t^{\mathcal{A}}(n)\right) \\
P\left[s'_t(n) = \mathcal{A} | s_t(n) = \mathcal{A}, \tau_t^{\mathcal{D}}(n) = 1\right] &= 1 - P^{\mathcal{D}}\left(\tau_t^{\mathcal{D}}(n)\right) \cdot \left(1 - \varphi_t^{\mathcal{A}}(n)\right) \\
P\left[s'_t(n) = \mathcal{D} | s_t(n) = 0, \tau_t^{\mathcal{D}}(n) = 1\right] &= P^{\mathcal{D}}\left(\tau_t^{\mathcal{D}}(n)\right) \\
P\left[s'_t(n) = 0 | s_t(n) = 0, \tau_t^{\mathcal{D}}(n) = 1\right] &= 1 - P^{\mathcal{D}}\left(\tau_t^{\mathcal{D}}(n)\right) \\
P\left[s'_t(n) = \mathcal{D} | s_t(n) = \mathcal{D}, \tau_t^{\mathcal{D}}(n) = 1\right] &= 1.
\end{aligned}
$$

We note that Protect moves are only feasible for the adversary if they already control the contested node. In the same manner we can compute the constituent probabilities for $s_{t+1}(n)$. These constituent probabilities are then:

$$
\begin{aligned}
P\left[s_{t+1}(n) = s'_t(n) | a_t^{\mathcal{A}}(n) = 0\right] &= 1 \\
P\left[s_{t+1}(n) = 0 | s'_{t+1}(n) = 0, a_t^{\mathcal{A}}(n)\right] &= 1 \\
P\left[s_{t+1}(n) = \mathcal{A} | s'_t(n') = \mathcal{D}, \tau_t^{\mathcal{A}}(n) = 1, \varphi_t^{\mathcal{D}}(n)\right] &= P^{\mathcal{A}}\left(\tau_t^{\mathcal{A}}(n)\right) \cdot \left(1 - \varphi_t^{\mathcal{D}}(n)\right) \\
P\left[s_{t+1}(n) = \mathcal{D} | s'_t(n) = \mathcal{D}, \tau_t^{\mathcal{A}}(n) = 1, \varphi_t^{\mathcal{D}}(n)\right] &= 1 - P^{\mathcal{A}}\left(\tau_t^{\mathcal{A}}(n)\right) \cdot \left(1 - \varphi_t^{\mathcal{D}}(n)\right).
\end{aligned}
$$

The transition probabilities for non-game nodes are considerably simpler. The probability of a take move by agent k on node n at epoch t, where n is a node that is accessible only to agent k, is a Bernoulli random variable with probability of a successful take $P^k\left(\tau_t^k(n)\right)$.

C.2 Updating the Belief Distribution

The defender selects action $\pi_{rob}^{\mathcal{D}}(R_t)$, where R_t is a random variable that is independent of all other random variables, has a state space identical to that of the state of the precedence and development process graph, has $\{P[R_t|\mathscr{I}_t^{\mathcal{D}}]\} = X_t^{\mathcal{D}}$ as its probability mass vector, and reveals its realization at epoch t. Note that when the detection system observes the state of

the current precedence graph precisely (i.e., $P[Z_t|S_t] = 1$ if and only if $z_t(n) = s_t(n)$ for all n), then the defender knows the state of the precedence graph exactly and this heuristic selects action $\pi^D(S_t)$ with probability 1.

The array $\{P[S_t|\mathscr{I}_t^{\mathcal{D}}]\}$ is called the *belief distribution*. We note that

$$\{Z_t, Z_{t-1}, \ldots, A_{t-1}^{\mathcal{D}}, A_{t-2}^{\mathcal{D}}, \ldots, \pi^{\mathcal{A}}, \pi^{\mathcal{D}}\}$$

can be deduced from

$$\{Z_t, Z_{t-1}, \ldots, R_{t-1}, R_{t-2}, \ldots, \pi^{\mathcal{A}}, \pi^{\mathcal{D}}\},$$

but not vice versa, where R_t is the realization of the random variable having the same state space as $\{S_t, t = 0, 1, \ldots\}$ and probability mass function $P\big[\,\cdot\,|\mathscr{I}_t^{\mathcal{D}}\big]$ such that $A_t^{\mathcal{D}} = \pi^{\mathcal{D}}(R_t)$ with probability $P\big[R_t|\mathscr{I}_t^{\mathcal{D}}\big]$. (With some potential for confusion, we are using the same terms to represent random variables and their realizations.) We remark that given R_t, we can determine $A_t^{\mathcal{D}} = \pi^{\mathcal{D}}(R_t)$; however, there may be several values of R_t such that $A_t^{\mathcal{D}} = \pi^{\mathcal{D}}(R_t)$. Thus, going forward we will assume $\mathscr{I}_t^{\mathcal{D}} = \{Z_t, Z_{t-1}, \ldots, R_{t-1}, R_{t-2}, \ldots, \pi^{\mathcal{A}}, \pi^{\mathcal{D}}\}$.

Noting $\mathscr{I}_{t+1}^{D} = \{Z_{t+1}^{\mathcal{D}}, R_t, \mathscr{I}_t^{\mathcal{D}}\}$,

$$\begin{aligned} P\left[S_{t+1}|\mathscr{I}_{t+1}^{\mathcal{D}}\right] &= P\big[S_{t+1}|Z_{t+1}, R_t, \mathscr{I}_t^{\mathcal{D}}\big] \\ &= \frac{P\big[S_{t+1}, Z_{t+1}, R_t|\mathscr{I}_t^{\mathcal{D}}\big]}{\sum\limits_{S_{t+1}} P\big[S_{t+1}, Z_{t+1}, R_t|\mathscr{I}_t^{\mathcal{D}}\big]} \\ &= \frac{P\big[S_{t+1}, Z_{t+1}, R_t|\mathscr{I}_t^{\mathcal{D}}\big]}{\sum\limits_{S_t}\sum\limits_{S_{t+1}} P\big[S_{t+1}, S_t, Z_{t+1}, R_t|\mathscr{I}_t^{\mathcal{D}}\big].} \end{aligned}$$

It is then straightforward to show that

$$P\big[S_{t+1}, S_t, Z_{t+1}, R_t|\mathscr{I}_t^{\mathcal{D}}\big] = P\big[Z_{t+1}|S_{t+1}\big]P\big[S_{t+1}|S_t, \pi^{\mathcal{A}}(S_t), \pi^{\mathcal{D}}(R_t)\big]P\big[S_t, R_t|\mathscr{I}_t^{\mathcal{D}}\big],$$

where $P\big[S_t, R_t | \mathscr{I}_t^{\mathcal{D}}\big] = P\big[R_t | S_t, \mathscr{I}_t^{\mathcal{D}}\big] P\big[S_t | \mathscr{I}_t^{\mathcal{D}}\big]$. By assumption,

$$P\big[R_t | S_t, \mathscr{I}_t^{\mathcal{D}}\big] = P\big[R_t | \mathscr{I}_t^{\mathcal{D}}\big],$$

and hence

$$P\big[S_{t+1}, S_t, Z_{t+1}, R_t | \mathscr{I}_t^{\mathcal{D}}\big] = P\big[Z_{t+1} | S_{t+1}\big] P\big[S_{t+1} | S_t, \pi^{\mathcal{A}}(S_t), \pi^{\mathcal{D}}(R_t)\big] P\big[R_t | \mathscr{I}_t^{\mathcal{D}}\big] P\big[S_t | \mathscr{I}_t^{\mathcal{D}}\big].$$

C.3 Thompson Sampling

The general, non-stationary Thompson sampling algorithm progresses as follows. Suppose the decision-maker (DM) has an action set $\mathscr{A}_t$ for $t = 0, \ldots, T$. After the DM chooses action a_t, the system generates an outcome y_t according to a known probability distribution $q_{\theta_t}(\cdot | a_t)$, where θ_t is unknown but learned through successive Bayesian updates. The system then generated a reward $r_t(y_t)$ according to a known reward function r_t. The Thompson sampling policy at each epoch t then samples the estimated parameter $\hat{\theta}_t$ according to the Bayesian belief distribution over possible θ_t, chooses the action that maximizes $\mathbb{E}_{q_{\theta_t}}[r(y_t) | a_t]$, and updates the belief distribution according to the observation y_t.

We see the connection between our heuristic and Thompson sampling by considering $\theta_t = S_t$, $y_t = R_t$, and $-r_t(R_t) = V_t^{\pi_{rob}^{\mathcal{D}}}(R_t | \pi_t^{\mathcal{A}})$ (the cost-to-go evaluation of $\pi_{rob}^{\mathcal{D}}$).

C.4 Computational Tractability

For the sake of simplicity of exposition in this section, we assume that none of the nodes in the graph "exist", in the sense of Section 4.3 in which they are in either state $\mathcal{A}$ or $\mathcal{D}$. Further, we restrict our attention to "take" moves, and denote the take action simply by $a^{\mathcal{A}}$ for the attacker, and $a^{\mathcal{D}}$ for the defender. Finally, we assume that the nodes states $\mathcal{A}$ and $\mathcal{D}$ are encoded such that $\mathcal{A} > \mathcal{D}$.

Due to the nature of *trust* problem applications, we have demonstrated that the *prece-*

dence graph structure is a natural model for the state of the system that facilitates conceptual domain understanding. Further, the information primitives (attack/counter-attack success probabilities) are provided at the level of nodes in the precedence graph. In order to utilize the POMG framework, we demonstrated how to map these information primitives into POMG model primitives in the form of conditional state transition probabilities, $\{P[S_{t+1}|S_t, A_t^{\mathcal{A}}, A_t^{\mathcal{D}}]\}$.

Since the information primitives, *i.e.* the attack/counter-attack success probabilities $\left\{P^{\mathcal{A}}\left(a^{\mathcal{A}}(n)\right)\right\}$ and $\left\{P^{\mathcal{D}}\left(a^{\mathcal{D}}(n)\right)\right\}$, differ in form to the POMG model primitives, *i.e.* the conditional state transition probabilities, $\{P[S_{t+1}|S_t, A_t^{\mathcal{A}}, A_t^{\mathcal{D}}]\}$, prior to any solution procedure we must go through a computationally taxing pre-processing step which takes these information primitives and computes all of the relevant conditional state transition probabilities. There are 2^N possible states and $2^N \times 2^N \times N^{\mathcal{A}} \times N^{\mathcal{D}}$ transition probabilities, where $N^{\mathcal{A}}$ is the number of possible node attacks available to the attacker, and $N^{\mathcal{D}}$ is likewise defined for the defender. Each conditional state transition probability requires computation of the conditional *node* transition probabilities $\{P[s_{t+1}(n)|S'_t, A_t^{\mathcal{A}}, A_t^{\mathcal{D}}]\}$ and $\{P[s'_t(n)|S_t, A_t^{\mathcal{A}}, A_t^{\mathcal{D}}]\}$. For large N, these computations become intractable.

In the following proposition, we present a result that demonstrates that the matrix of conditional state transition probabilities is sparse due to the nature of the rules governing state dynamics. We use this fact to significantly reduce the computational burden of this pre-processing step in our numerical results in Section 4.6.

Let $f(s, A^{\mathcal{D}})$ be the N-dimensional vector representing the best possible outcome for the defender, in which all of the defender's counter-attacks are successful. Likewise, let $g(s, A^{\mathcal{A}})$ be the N-dimensional vector representing the worst possible outcome for the defender in which all of the defender's counter-attacks are unsuccessful and all of the at-

tacker's attacks are successful. These vectors have the following closed form expressions:

$$f(s, A^{\mathcal{D}})[n] = \begin{cases} \mathcal{D}, \text{ if } P^{\mathcal{D}}\big(a^{\mathcal{D}}(n)\big) > 0, a^{\mathcal{D}}(n) > 0, s(n) = \mathcal{A} \\ \mathcal{D}, \text{ if } s(n) = \mathcal{D} \\ \mathcal{A}, \text{ if } s(n) = \mathcal{A}, a^{\mathcal{D}}(n) = 0 \end{cases}$$

$$g(s, A^{\mathcal{A}})[n] = \begin{cases} \mathcal{A}, \text{ if } P^{\mathcal{A}}\big(a^{\mathcal{A}}(n)\big) > 0, a^{\mathcal{A}}(n) > 0 \\ \mathcal{A}, \text{ if } s(n) = \mathcal{A} \\ \mathcal{D}, \text{ if } s(n) = \mathcal{D}, a^{\mathcal{A}}(n) = 0. \end{cases}$$

Finally, let $\preceq$ be the binary operator, which induces a partial order on the state space, such that $s \preceq s'$ if and only if $s[n] \leq s'[n]$ for all $n \in N$.

Proposition 26. *Given* $(\pi^{\mathcal{A}}, \pi^{\mathcal{D}})$*,*

$$f\big(S_t, \pi^{\mathcal{D}}(S_t)\big) \preceq S_{t+1} \preceq g\big(S_t, \pi^{\mathcal{A}}(S_t)\big).$$

Proof. This is a straightforward application of the state dynamics rules. The defender cannot do better than all of his counter-attacks being successful, and cannot do worse than all of his counter-attacks being unsuccessful and all of the attacker's attacks being successful. □

Let $s : s' \triangleq \{s'' \in \{\mathcal{D}, \mathcal{A}\}^N : s \preceq s'' \preceq s'\}$, the set of all state vectors "between" s and s' in the sense of the partial ordering induced by $\preceq$. The following result is a straightforward corollary to the prior proposition that demonstrates the sparsity of the matrix of conditional state transition probabilities.

Corollary 6. $P[S_{t+1} = s | S_t, A_t^{\mathcal{A}}, A_t^{\mathcal{D}}] = 0$*, for all* $s \notin f(S_t, A_t^{\mathcal{D}}) : g(S_t, A_t^{\mathcal{A}})$.

REFERENCES

[1] S. C. Albright, "Structural Results for Partially Observable Markov Decision Processes," *Operations Research*, vol. 27, pp. 1041–1053, 1979.

[2] Y. Aviv and A. Pazgal, "A Partially Observed Markov Decision Process for Dynamic Pricing," *Management Science*, vol. 51(9), pp. 1400–1416, 2005.

[3] J. L. Bander and C. C. White III, "Markov Decision processes with noise-corrupted and delayed state observations," *Journal of the Operational Research Society*, vol. 50, pp. 660–668, 1999.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996, vol. 5.

[5] R. R. Bishop and C. C. White III, "Sequential Decision Making Affected by Partially Observed Exogenous Forces," *In Review*, 2019.

[6] J. Bismut, "An Introductory Approach to Duality in Optimal Stochastic Control," *SIAM Review*, vol. 20(1), pp. 62–78, 1978.

[7] D. Brown, J. Smith, and P. Sun, "Information Relaxations and Duality in Sotchastic Dynamic Programs," *Operations Research*, vol. 58(4), pp. 785–801, 2010.

[8] Y. Chang, A. L. Erera, and C. C. White III, "Partially Observed Multi-objective Markov Games," *Annals of Operations Research*, vol. 235(1), pp. 103–128, 2015.

[9] ——, "Value of Information for a Leader-Follower Partially Observed Markov Game," *Annals of Operations Research*, vol. 235(1), pp. 129–153, 2015.

[10] ——, "Risk Assessment of Deliberate Contamination of Food Production Facilities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47(3), pp. 381–393, 2017.

[11] Y.-C. Chen, D. Campbell, V. Mooney, S. Grijalva, B. Eames, A. Outkin, E. Vugrin, R. Helinski, and B. Anthony, "Power Grid Bad Data Injection Attack Modeling in PRESTIGE," *GOMACTech, Albuquerque NM*, 2019.

[12] Y.-C. Chen, T. Gieseking, D. Campbell, V. Mooney, and S. Grijalva, "A Hybrid Attack Model for Cyber-Physical Security Assessment in Electricity Grid," *IEEE Texas Power and Energy Conference (TPEC), College Station, TX*, pp. 1–6, 2019.

[13] J.-K. Chong, T.-H. Ho, and C. Camerer, "A Generalized Cognitive Hierarchy Model of Games," *Games and Economic Behavior*, vol. 99, pp. 257–274, 2016.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[15] B. K. Eames, A. V. Outkin, S. Walsh, J. R. Mayo, J. R. Hamlet, J. M. Eldridge, R. C. Armstrong, M. P. Napier, G. D. Wyss, E. D. Vugrin, and M. L. Holmes, "Fundamental Trust Analysis," Sandia National Laboratories, Tech. Rep., 2016.

[16] Y. Ephraim and N. Merhav, "Hidden Markov Processes," *IEEE Transactions on Information Theory*, vol. 48(6), pp. 1518–1569, 2002.

[17] K. J. Ferreira, D. Simchi-Levi, and H. Wang, "Online Network Revenue Management Using Thompson Sampling," *Operations Research*, vol. 66(6), pp. 1586–1602, 2018.

[18] M. Galiardi, E. Vugrin, B. Eames, A. Outkin, G. Wyss, J. Hamlet, R. Helinski, B. Anthony, M. Napier, J. Eldridge, A. Bertels, and M. Holmes, "On Modeling Detection for Quantitative Trust Analysis," *GOMACTech 2018, Miami FL*, 2018.

[19] S. Gavirneni, R. Kapuscinski, and S. Tayur, "Value of Information in Capacitated Supply Chains," *Management Science*, vol. 45(1), pp. 16–24, 1999.

[20] H. Goldstein, *Multilevel Statistical Models, 4th Edition*. John Wiley & Sons, Ltd., 2011.

[21] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[22] M. Hauskrecht, "Value-Function Approximations for Partially Observable Markov Decision Processes," *Journal of Artificial Intelligence Research*, vol. 13, pp. 33–94, 2000.

[23] J. H. Holland, "Adaptation in Natural and Artificial Systems," *University of Michigan Press*, 1975.

[24] S. Jones, A Outkin, J. Gearhart, J. Hobbs, C. Siirola, S. Phillips, D. Verzi, D. Tauritz, S. Mulder, and A. Naugl, "Evaluating Moving Target Defense with PLADD," *Sandia National Laboratories*, 2015.

[25] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains," *Artificial Intelligence*, vol. 101(1-2), pp. 99–134, 1998.

[26] KPMG, "Cost of Capital Study 2018: New Business Models - Risks or Rewards," Tech. Rep., 2018.

[27] H. L. Lee, K. C. So, and C. S. Tang, "The Value of Information Sharing in a Two-Level Supply Chain," *Management Science*, vol. 46(5), pp. 626–643, 2000.

[28] Q. Li and P. Yu, "Multimodularity and Its Applications in Three Stochastic Dynamic Inventory Problems," *Manufacturing & Service Operations Management*, vol. 16(3), pp. 455–463, 2014.

[29] W. S. Lovejoy, "Some Monotonicity Results for Partially Observed Markov Decision Processes," *Operations Research*, vol. 35(5), pp. 736–743, 1987.

[30] ——, "Computationally Feasible Bounds for Partially Observed Markov Decision Processes," *Operations Research*, vol. 39(1), pp. 162–175, 1991.

[31] S. Malladi, A. Erera, and C. C. White III, "A Partially Observed Inventory Control Problem," *In Review*, 2018.

[32] K. Murota, *Discrete convex analysis*. SIAM, 2003.

[33] A. V. Outkin, B. K. Eames, M. A. Galliardi, S. Walsh, E. D. Vugrin, B. Heerskin, J. Hobbs, and G. D. Wyss, "GPLADD: Quantifying Trust in Government and Commercial Systems A Game-Theoretic Approach," *ACM Transactions on Privacy and Security*, vol. 22 (3), 2019.

[34] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: An anytime algorithm for POMDPs," *Proceedings of International Joint Conference on Artificial Intelligence*, 2003.

[35] E. L. Porteus, "On the Optimality of Structured Policies in Countable Stage Decision Processes," *Management Science*, vol. 22(2), pp. 148–157, 1975.

[36] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, Inc., 2011.

[37] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken: John Wiley & Sons, 2010.

[38] L. Rogers, "Pathwise Stochastic Optimal Control," *SIAM Journal on Control and Optimization*, vol. 46(3), pp. 1116–1132, 2007.

[39] D. J. Russo and B. Van Roy, "Learning to Optimize via Posterior Sampling," *Mathematics of Operations Research*, vol. 39(4), pp. 1221–1243, 2014.

[40] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A Tutorial on Thompson Sampling," *Foundation and Trends in Machine Learning*, vol. 11(1), pp. 1–96, 2018.

[41] B. Sandikci, L. M. Maillart, A. J. Schaefer, O. Alagoz, and M. S. Roberts, "Estimating the Patient's Price of Privacy in Liver Transplantation," *Operations Research*, vol. 56(6), pp. 1393–1410, 2008.

[42] B. Sandikci, L. M. Maillart, A. J. Schaefer, and M. S. Roberts, "Alleviating the Patient's Price of Privacy Through a Partially Observable Waiting List," *Management Science*, vol. 59(8), pp. 1836–1854, 2013.

[43] R. F. Serfozo, "Monotone Optimal Policies for Markov Decision Processes," *Mathematical Programming Studies Stochastic Systems: Modeling, Identification and Optimization*, vol. 11, pp. 202–215, 1976.

[44] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassibis, "A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-play," *Science*, vol. 362(6419), pp. 1140–11 144, 2018.

[45] R. D. Smallwood and E. J. Sondik, "The Optimal Control of Partially Observable Markov Processes over a Finite Horizon," *Operations Research*, vol. 21(5), pp. 1071–1088, 1973.

[46] J. E. Smith and K. F. McCardle, "Structural Properties of Stochastic Dynamic Programs," *Operations Research*, vol. 50(5), pp. 796–809, 2002.

[47] M. Sobel, "Myopic Solutions of Markov Decision Processes and Stochastic Games," *Operations Research*, vol. 29, pp. 995–1009, 1981.

[48] E. J. Sondik, "The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs," *Operations Research*, vol. 26(2), pp. 282–304, 1978.

[49] M. T. Spaan and N. Vlaasis, "Perseus: Randomized Point-based Value Iteration for POMDPs," *Journal of Artificial Intelligence Research*, vol. 24, pp. 195–220, 2005.

[50] D. O. Stahl and P. W. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, vol. 10(1), pp. 218–254, 1995.

[51] L. N. Steimle, D. L. Kaufman, and B. T. Denton, "Multi-model Markov Decision Processes," *In Review*, 2018.

[52] W. R. Thompson, "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, vol. 25(3), pp. 285–294, 1933.

[53] D. M. Topkis, "Minimizing a Submodular Function on a Lattice," *Operations Research*, vol. 26, pp. 305–321, 1978.

[54] J. Treharne and C. Sox, "Adaptive Inventory Control for Nonstationary Demand and Partial Information," *Management Science*, vol. 48(5), pp. 607–624, 2002.

[55] T. G. Tryphon and A. Lindquist, "The Separation Principle in Stochastic Control, Redux," *IEEE Transactions on Automatic Control*, vol. 58(10), pp. 2481–2494, 2013.

[56] A. F. Veinott Jr., "Optimal Policy for a Multi-product, Dynamic, Nonstationary Inventory Problem," *Management Science*, vol. 12(3), pp. 206–222, 1965.

[57] ——, "Optimal Policy in a Dynamic, Single product, Nonstationary Inventory Model with Several Demand Classes," *Operations Research*, vol. 21(5), pp. 1071–1078, 1965.

[58] C. C. White III, "A Markov Quality Control Process Subject to Partial Observation," *Management Science*, vol. 23, pp. 843–852, 1977.

[59] ——, "Optimal Control-limit Strategies for a Partially Observed Replacement Problem," *Int. J. Syst. Sci*, vol. 10, pp. 321–331, 1979.

[60] ——, "Monotone Control Laws for Noisy, Countable-state Markov Chains," *Eur. J. Opns. Res.*, vol. 5, pp. 124–132, 1980.

[61] ——, "Structured Policy Results for a Single Stage Decisionmaking Under Uncertainty," *IEEE Trans. Syst. Man Cybernet*, vol. 10, pp. 891–894, 1980.

[62] C. C. White III and D. P. Harrington, "Application of Jensen's inequality to adaptive suboptimal design," *Journal of Optimization Theory and Applications*, vol. 32(1), pp. 89–99, 1980.

[63] D. J. White, "Monotone Value Iteration for Discounted Finite Markov Decision Processes," *Journal of Mathematical Analysis and Applications*, vol. 109(2), pp. 311–324, 1985.

[64] P. Zipkin, "On the Structure of Lost-Sales Inventory Models," *Operations Research*, vol. 56, pp. 937–944, 2008.